Appl. Statist. (1998)
47, *Part* 3, *pp.* 299–350

Model-based geostatistics

P. J. Diggle† and J. A. Tawn

Lancaster University, UK

and R. A. Moyeed

Plymouth University, UK

[*Read before* The Royal Statistical Society *on Wednesday, November 12th, 1997, the President*, Professor R. N. Curnow, *in the Chair*]

Summary. Conventional geostatistical methodology solves the problem of predicting the realized value of a linear functional of a Gaussian spatial stochastic process $S(\mathbf{x})$ based on observations $Y_i = S(\mathbf{x}_i) + Z_i$ at sampling locations $\mathbf{x}_i$, where the Z_i are mutually independent, zero-mean Gaussian random variables. We describe two spatial applications for which Gaussian distributional assumptions are clearly inappropriate. The first concerns the assessment of residual contamination from nuclear weapons testing on a South Pacific island, in which the sampling method generates spatially indexed Poisson counts conditional on an unobserved spatially varying intensity of radioactivity; we conclude that a conventional geostatistical analysis oversmooths the data and underestimates the spatial extremes of the intensity. The second application provides a description of spatial variation in the risk of campylobacter infections relative to other enteric infections in part of north Lancashire and south Cumbria. For this application, we treat the data as binomial counts at unit postcode locations, conditionally on an unobserved relative risk surface which we estimate. The theoretical framework for our extension of geostatistical methods is that, conditionally on the unobserved process $S(\mathbf{x})$, observations at sample locations $\mathbf{x}_i$ form a generalized linear model with the corresponding values of $S(\mathbf{x}_i)$ appearing as an offset term in the linear predictor. We use a Bayesian inferential framework, implemented via the Markov chain Monte Carlo method, to solve the prediction problem for non-linear functionals of $S(\mathbf{x})$, making a proper allowance for the uncertainty in the estimation of any model parameters.

Keywords: Generalized linear mixed model; Geostatistics; Kriging; Markov chain Monte Carlo method; Spatial prediction

1. Introduction

The name *kriging* refers to a widely used method for interpolating or smoothing spatial data. Given a set of data y_i, $i = 1, \ldots, n$, at spatial locations $\mathbf{x}_i$, the kriging predictor for the underlying spatial surface, $S(\mathbf{x})$ say, takes the form

$$\hat{S}(\mathbf{x}) = \sum_{i=1}^{n} w_i(\mathbf{x}) y_i, \tag{1}$$

where the *kriging weights* $w_i(\mathbf{x})$ are derived from the estimated mean and covariance structure of the data. For a model-based derivation, we can assume that the data are generated by the model

†*Address for correspondence*: Department of Mathematics and Statistics, Lancaster University, Lancaster, LA1 4YF, UK.
E-mail: p.diggle@lancaster.ac.uk

© 1998 Royal Statistical Society 0035–9254/98/47299

$$Y_i = \mu + S(\mathbf{x}_i) + Z_i, \qquad i = 1, \ldots, n,$$

where μ is a constant mean effect, $S(\mathbf{x})$ is a stationary Gaussian process with $E[S(\mathbf{x})] = 0$ and $\text{cov}\{S(\mathbf{x}), S(\mathbf{x}')\} = \sigma^2\rho(\mathbf{x} - \mathbf{x}')$, and the Z_i are mutually independent $N(0, \tau^2)$. An equivalent formulation is that, conditionally on $S(\cdot)$, the Y_i are mutually independent, with

$$Y_i|S(\mathbf{x}_i) \sim N\{\mu + S(\mathbf{x}_i), \tau^2\}. \tag{2}$$

In the more applied statistical literature on kriging, these distributional assumptions are often not made explicitly. However, the *linear* predictor (1) might be regarded as a natural choice under Gaussian assumptions, since it then minimizes $E[\{\hat{S}(\mathbf{x}) - S(\mathbf{x})\}^2]$.

The bald statement that kriging is linear prediction conceals a large body of methodology, collectively known as *geostatistics* in acknowledgement of its origins in mineral exploration. Much of the early development of geostatistical methodology was undertaken by G. Matheron and colleagues at Fontainebleau, France. See, for example, Matheron (1970). More recent text-book accounts include Journel and Huijbregts (1978) and Isaaks and Srivastava (1989). Parallel independent developments in stochastic process prediction (Whittle, 1963) and in the analysis of spatial variation (Matérn, 1960) eventually led to the placing of geostatistical methods within the wider setting of spatial statistics. See, for example, Ripley (1981), chapter 4, or Cressie (1991), chapters 2–5.

Our aim in this paper is to extend the geostatistical method to situations in which the stochastic variation in the data is known to be non-Gaussian. In current geostatistical practice, the most widely implemented methodology for coping with non-Gaussian problems is trans-Gaussian kriging (Cressie (1991), pages 137–138) which consists of applying standard Gaussian methods after a marginal non-linear transformation, i.e. analyse transformed data $\phi(Y_i)$ for some specified function $\phi(\cdot)$. In contrast, our proposal is to embed linear kriging methodology within a more general distributional framework, analogous to the embedding of the Gaussian linear model for mutually independent data within the framework of the generalized linear model (Nelder and Wedderburn, 1972; McCullagh and Nelder, 1989).

The work has been motivated by the following two examples.

1.1. Example 1: radionuclide concentrations on Rongelap Island

Rongelap Island forms part of the Republic of the Marshall Islands and is located in the Pacific Ocean approximately 2500 miles south-west of Hawaii. The island experienced contamination due to fall-out from the Bikini Atoll nuclear weapons testing programme during the 1950s, and the former inhabitants of the island have been living in self-imposed exile on the much smaller island of Mejatto since 1985.

As part of a wider investigation to establish whether Rongelap can safely be resettled, the Marshall Islands National Radiological Survey has examined the current levels of ^{137}Cs contamination by *in situ* γ-ray counting at a set of $n = 157$ locations over the island. The γ-spectrometer records cumulative numbers of photon emissions in discrete energy ranges. Each radionuclide emits photons at specific energy levels, called *channels*, which, in principle, make the separate identification of each possible. In practice, some leakage of energy between channels occurs and to estimate the level of activity in the energy range corresponding to a particular radionuclide it is standard practice to subtract from the gross counts an estimate of the leakage, based on the (typically much smaller) average counts in adjacent energy ranges. According to the well-established theory of radioactive emissions, the counts Y_i at the n locations can therefore be treated approximately as realizations of mutually independent

Poisson random variables with expectations $M_i = t_i\,\lambda(\mathbf{x}_i)$, where t_i denotes the length of time over which the counts are recorded and $\lambda(\mathbf{x})$ measures the ^{137}Cs radioactivity at location $\mathbf{x}$. The approximation arises because of the effect of the leakage correction. For further details, see Diggle *et al.* (1997).

One objective of the Rongelap survey is to estimate $\lambda(\mathbf{x})$. Interest is also in non-linear functions of $\lambda(\mathbf{x})$ such as $\max\{\lambda(\mathbf{x})\}$, the location associated with this maximum, and with the regions of the island where radioactivity is above a specified threshold. To reflect the effective dose received by an individual utilizing a finite area around their home, there is also practical interest in these characteristics for the spatially averaged process $T_\psi(\mathbf{x})$ where, for any given positive value ψ,

$$T_\psi(\mathbf{x}) = \int_{\|\mathbf{u}\| \leqslant \psi} \lambda(\mathbf{x} - \mathbf{u})\,\mathrm{d}\mathbf{u}.$$

In the absence of any physically based model for a spatial trend in ^{137}Cs concentrations, it may be reasonable to assume that $\log\{\lambda(\mathbf{x})\} = \mu + S(\mathbf{x})$, where $S(\cdot)$ is a zero-mean, stationary Gaussian process and the parameter μ represents the mean log-intensity over the island. In this example, the Poisson distribution is clearly more appropriate than the Gaussian as a model for $Y_i|S(\mathbf{x}_i)$.

1.2. Example 2: campylobacter infections in north Lancashire and south Cumbria

The incidence of campylobacter infections in the UK has grown dramatically over the past 20 years, to the point where it is now the most common cause of enteric infection. Furthermore, the incidence is especially high around Lancaster and there is a suggestion of small scale spatial variation in relative risk (Jones and Telford, 1991). At Lancaster Royal Infirmary, Dr D. Telford has assembled a data set giving the spatial locations (as identified by residential unit postcodes) of all reported cases of enteric infections in the north Lancashire–south Cumbria region between 1991 and 1994. Using these data, it is possible to examine the spatial variation in the relative risk of campylobacter among all recorded cases of the three most common enteric infections (campylobacter, salmonella and cryptosporidia). The data which we shall analyse consist of the numbers Y_i of campylobacters and the total numbers m_i of recorded cases, at each of $n = 248$ unit postcode locations within postcode sectors LA8, LA9, LA10, LA21, LA22 and LA23. Assuming independent infections (for further discussion of which, see Section 6.3 later), it is reasonable to model the Y_i as conditionally independent binomial random variables, given the underlying spatial surface of relative risk. As in example 1, we shall use a stationary Gaussian process $S(\cdot)$ as the basis of an empirical model for the spatial variation in the probability, $P(\mathbf{x})$ say, that a case at $\mathbf{x}$ is a campylobacter infection, but using a logit transformation to map the domain of $S(\cdot)$ onto the unit interval; thus

$$\log[P(\mathbf{x})/\{1 - P(\mathbf{x})\}] = \mu + S(\mathbf{x}).$$

In both of our motivating examples, and for the standard Gaussian kriging model (2), the regression function $E[Y_i|S(\mathbf{x}_i)]$ varies spatially only through the value of $S(\mathbf{x})$ at location $\mathbf{x}_i$. Here, $S(\mathbf{x})$ can be thought of as being a surrogate for any spatial variation in unavailable explanatory variables for the observed $\mathbf{Y} = (Y_1, \ldots, Y_n)$. In some applications, there may be grounds for linking the spatial variation in the regression function to a vector of observable spatial explanatory variables, $\mathbf{d}(\mathbf{x})$ say. This leads to a more general formulation in which the spatial regression function takes the form

$$E[Y_i|S(\mathbf{x}_i)] = M_i = M(\mathbf{x}_i),$$

where

$$h\{M(\mathbf{x})\} = \mathbf{d}(\mathbf{x})^{\mathrm{T}}\boldsymbol{\beta} + S(\mathbf{x})$$

for a known link function $h(\cdot)$ and unknown parameters $\boldsymbol{\beta}$. In other words, conditionally on the Gaussian process $S(\mathbf{x})$, the data Y_i, $i = 1, \ldots, n$, follow a classical generalized linear model. Models of this kind are examples of generalized linear mixed models (Breslow and Clayton, 1993). In the present context, the role of the Gaussian process $S(\mathbf{x})$ is to explain the residual spatial variation after accounting for all known explanatory variables. In the linear Gaussian setting, the inclusion of spatially varying explanatory variables is known as *universal kriging*.

In particular applications, the inferential focus may be on the regression parameters $\boldsymbol{\beta}$, on properties of $S(\cdot)$ or on the conditional distribution of $S(\cdot)$ given the data $\mathbf{Y}$. In standard kriging, the primary objective is to predict the realization of the random function $M(\mathbf{x})$ given $\mathbf{Y}$, and the model parameters are of limited interest in themselves. This is also the focus in both of our motivating examples, except that in example 1 there is also a specific interest in predicting non-linear functionals of $M(\cdot)$, e.g. the maximum value of $M(\mathbf{x})$ over the island, or those parts of the island for which $M(\mathbf{x})$ exceeds a given threshold.

When the regression parameters $\boldsymbol{\beta}$ are of direct interest, it is important to remember that these parameters have a conditional rather than a marginal interpretation. In particular, $E[Y_i|S(\mathbf{x}_i)]$ and $E[Y_i]$ differ in their structural dependence on the explanatory variables $\mathbf{d}(\mathbf{x}_i)$, so the interpretation of $\boldsymbol{\beta}$ requires care. Only in the case where $Y_i|S(\mathbf{x}_i)$ is Gaussian and the link function is the identity can $\boldsymbol{\beta}$ also be treated as the regression parameter for the marginal regression function $E[Y_i]$.

We shall adopt a Bayesian framework for inference and prediction, implemented using Markov chain Monte Carlo (MCMC) methods (Smith and Roberts, 1993). This enables us to incorporate the uncertainty due to estimation of parameters in both the systematic and the stochastic components of the model into the reported precision of our results. It turns out that the effect of parameter uncertainty can be substantial, especially when the objective is to predict a non-linear functional of the underlying spatial process. Le and Zidek (1992) and Handcock and Stein (1993) also proposed a Bayesian formulation of the problem, but in the context of the linear Gaussian model (2).

The combination of generalized linear mixed modelling and MCMC sampling for spatial data has a rapidly growing literature. A major area of application, following early work by Clayton and Kaldor (1987) and Besag *et al.* (1991), is to disease mapping problems. Here, the basic data take the form of the number of cases of a particular disease and the corresponding population size in each of a number of discrete spatial regions. Spatial smoothness is built into the analysis by using a Markov random field model to describe the dependence in disease rates between spatially adjacent regions. Subsequent work in this area is reviewed in Mollié (1996). In the discussion of Besag *et al.* (1991), Raftery and Banfield (1991) suggested that in some applications it might be more appropriate to model the spatial dependence continuously by analogy with classical geostatistics, rather than among a discrete set of points or regions. Cressie (1994) made a similar suggestion in the discussion of Handcock and Wallis (1994). Freulon (1994) described the use of the Gibbs sampler for simulating spatial processes consisting of a latent Gaussian process with observations subject to noise, including Poisson noise as a special case. For models of this form, Lawson *et al.* (1996) presented results using both full Bayesian inference and a quadratic approximation to the log-likelihood as a

computationally easy alternative. Almost all this work assumes a Gaussian model for the underlying spatial variation. An exception is Wolpert and Ickstadt (1997), who considered Cox point processes with a gamma random field model for the spatial variation in the local intensity of points.

The remainder of the paper is structured as follows. Section 2 discusses standard kriging procedures from a statistical modelling perspective. Section 3 describes a more general formulation based on the generalized linear mixed model. Section 4 describes an implementation of MCMC methodology to solve the associated inferential problems. Section 5 considers the extension of the variogram, which is a standard tool in conventional geostatistics, to the more general setting. Section 6 presents applications to a simulated case-study and to our two motivating examples. The paper closes with a discussion of some open questions.

2. Kriging: existing methods

Let $S \equiv \{S(\mathbf{x}): \mathbf{x} \in \mathbb{R}^p\}$ denote a stochastic process, called the *signal*, whose realized values are not directly observable. Let $\mathbf{Y} \equiv (Y_1, \ldots, Y_n)$ denote a random vector which is stochastically dependent on S, and whose realized values *are* directly observable. We shall think of Y_i as a 'noisy' version of $S(\mathbf{x}_i)$ for a prescribed set of locations $\mathbf{x}_i$, $i = 1, \ldots, n$, and consider the problem of predicting the realized values of functionals of S from the data, y_i, $i = 1, \ldots, n$.

Let T be any functional of S. We define the *kriging predictor* of T as that function, $\hat{T} \equiv \hat{T}(\mathbf{Y})$ say, which minimizes the prediction mean-square error, $E[(T - \hat{T})^2]$. From a well-known result we have that the kriging predictor for T is $\hat{T} = E[T|\mathbf{Y}]$, which gives a point prediction of T. We call $\mathrm{var}(T|\mathbf{Y})$ the *prediction variance* for T.

2.1. Linear kriging

Implementation of the kriging predictor is extremely straightforward under the following assumptions:

(a) the process S is Gaussian, with mean $E[S(\mathbf{x})] = 0$ and covariance function $\gamma(\mathbf{x}, \mathbf{x}') = \mathrm{cov}\{S(\mathbf{x}), S(\mathbf{x}')\} = \sigma^2 \rho(\mathbf{x} - \mathbf{x}')$;
(b) conditionally on S, the Y_i, $i = 1, \ldots, n$, are mutually independent Gaussian random variables with expectations $\mu(\mathbf{x}_i) + S(\mathbf{x}_i)$ and common variance τ^2;
(c) the target functional T is a linear functional of S.

Let $\boldsymbol{\mu}$ denote the n-element vector with ith element $\mu(\mathbf{x}_i)$, $\mathbf{g}(\mathbf{x})$ the n-element vector with ith element $\gamma(\mathbf{x}_i, \mathbf{x})$, G the $n \times n$ matrix with (i, j)th element $\gamma(\mathbf{x}_i, \mathbf{x}_j)$ and I the $n \times n$ identity matrix. Then, using standard properties of the multivariate normal distribution (e.g. Mardia *et al.* (1979), p. 63) it follows that the kriging predictor for $S(\mathbf{x})$ is

$$\hat{S}(\mathbf{x}) = \mathbf{g}(\mathbf{x})^{\mathrm{T}}(\tau^2 I + G)^{-1}(\mathbf{Y} - \boldsymbol{\mu}), \tag{3}$$

and the prediction variance, $V(\mathbf{x}) = \mathrm{var}\{S(\mathbf{x})|\mathbf{Y}\}$, is

$$V(\mathbf{x}) = \sigma^2 - g(\mathbf{x})^{\mathrm{T}}(\tau^2 I + G)^{-1} g(\mathbf{x}). \tag{4}$$

Equation (3) gives the explicit form of the *linear* kriging predictor (1).

Now consider prediction for the linear functional

$$T = \int_A k(\mathbf{x})\, S(\mathbf{x})\, \mathrm{d}\mathbf{x},$$

where $k(\cdot)$ is any known function and A is some region of interest. Using the linearity of the expectation operator, it follows immediately that the kriging predictor of T is

$$\hat{T} = \int_A k(\mathbf{x})\, \hat{S}(\mathbf{x})\, \mathrm{d}\mathbf{x}, \tag{5}$$

with $\hat{S}$ given by equation (3), and similarly that the associated prediction variance is

$$V_T = \mathrm{var}(T|\mathbf{Y}) = \int_A \int_A k(\mathbf{x})\, V(\mathbf{x}, \mathbf{x}')\, k(\mathbf{x}')\, \mathrm{d}\mathbf{x}\, \mathrm{d}\mathbf{x}',$$

where

$$V(\mathbf{x}, \mathbf{x}') = \gamma(\mathbf{x}, \mathbf{x}') - g(\mathbf{x})^{\mathrm{T}}(\tau^2 I + G)^{-1} g(\mathbf{x}')$$

corresponds to $\mathrm{cov}\{S(\mathbf{x}), S(\mathbf{x}')|\mathbf{Y}\}$.

An approximate 95% prediction interval is $\hat{T} \pm 2V_T^{1/2}$. Although the kriging predictor can be derived as the optimal linear predictor without explicit reference to a Gaussian model, linear prediction is only compelling under Gaussian assumptions.

2.2. Disjunctive kriging

In the geostatistical literature there are also methods which use non-linear functions of the data to approximate functionals of S. One such method is disjunctive kriging (Matheron, 1976; Armstrong and Matheron, 1986a, b). Disjunctive kriging seeks to obtain an optimal predictor among the class of all linear combinations of univariate functions of the data. In particular, within this class, the minimum mean-squared error predictor of a functional of S at any given location $\mathbf{x}$, $\hat{T}\{S(\mathbf{x})\}$, takes the form

$$\hat{T}\{S(\mathbf{x})\} = \sum_{i=1}^{n} u_i\{Y(\mathbf{x}_i)\}, \tag{6}$$

where $\{u_i,\ i = 1, \ldots, n\}$ are measurable square integrable functions satisfying the following disjunctive kriging equations:

$$E[T\{S(\mathbf{x})\}|Y(\mathbf{x}_j)] = \sum_{i=1}^{n} E[u_i\{Y(\mathbf{x}_i)\}|Y(\mathbf{x}_j)], \qquad \text{for } j = 1, \ldots, n. \tag{7}$$

The solution to equations (7) assumes knowledge of only bivariate distributions. Note also that, if an arbitrary non-Gaussian bivariate distribution is used to evaluate the conditional expectations in equations (7), this begs the question of whether the assumed bivariate distributions are compatible with any underlying spatial process. For further discussion, see Armstrong and Matheron (1986a, b).

2.3. Indicator kriging

Another method which is used for deriving a predictor of the form (6) is indicator kriging (Journel, 1983). As the name suggests, indicator kriging is basically the application of kriging to indicator transformations of the data. Additionally, this method makes no parametric assumptions about the underlying bivariate distributions. Although the performance of

indicator kriging is relatively similar to other kriging methods, it is quite cumbersome to implement in practice and it has a rather weak theoretical foundation; see Papritz and Moyeed (1997).

2.4. The variogram

Typically, any unknown parameters involved in the specification of the covariance structure $\gamma(\mathbf{x}, \mathbf{x}')$ and the mean function $\mu(\mathbf{x})$ are estimated by virtually *ad hoc* methods. A statistic that is widely used for estimating the covariance structure is the variogram, which is defined as

$$C(\mathbf{u}) = \tfrac{1}{2} \operatorname{var}[\{Y(\mathbf{x} + \mathbf{u}) - Y(\mathbf{x})\}], \tag{8}$$

for any random process $\{Y(\mathbf{x}): \mathbf{x} \in R^2\}$. When $\mu(\mathbf{x}) = \mu$, a constant, a nonparametric estimator of the variogram is the empirical variogram $\tilde{C}(\mathbf{u})$, given by

$$\tilde{C}(\mathbf{u}) = \frac{1}{2|N(\mathbf{u})|} \sum \{Y(\mathbf{x}_i) - Y(\mathbf{x}_j)\}^2, \tag{9}$$

where the sum is over $N(\mathbf{u}) \equiv \{(i, j): \mathbf{x}_i - \mathbf{x}_j = \mathbf{u}\}$ and $|N(\mathbf{u})|$ is the number of distinct elements of $N(\mathbf{u})$. In practice, when the spatial distribution of the data locations $\mathbf{x}_i$ is irregular, equation (9) is often calculated after grouping the spatial separations $\mathbf{u}$ into discrete classes. Also, when $\mu(\mathbf{x})$ is not constant, equation (9) is applied to residuals after subtracting preliminary estimates of $\mu(\mathbf{x}_i)$ from the Y_i.

2.5. Parametric estimation of covariance structure

Common geostatistical practice is to fit a parametric model for the covariance structure by a direct comparison between the empirical and theoretical variograms, using curve fitting methods; see, for example, Cressie (1991). However, a growing body of work on inference for spatial stochastic processes has resulted in a move towards the adoption of likelihood-based methods of parameter estimation. See, for example, Mardia and Marshall (1984), Warnes and Ripley (1987), Vecchia (1988, 1992), Mardia and Watkins (1989), Zimmerman (1989) and Laslett (1994). A potentially serious limitation of conventional geostatistical methodology is that it uses the prediction variance as an estimate of precision without taking account of the uncertainty involved in estimating the parameters of the assumed covariance structure of S. As noted earlier, two exceptions are Le and Zidek (1992) and Handcock and Stein (1993), which address the question of parameter uncertainty by adopting a Bayesian formulation of the problem.

3. Generalized linear prediction

The distinctive feature of kriging methodology derived from the Gaussian assumptions (a) and (b) in Section 2 is that the predictor $\hat{S}(\mathbf{x})$ is linear in the data $\mathbf{Y}$. We now extend this methodology in exactly the same way that generalized linear models (Nelder and Wedderburn, 1972; McCullagh and Nelder, 1989) extend the classical Gaussian linear model. In the following model we assume the following:

(a) S is a stationary Gaussian process with $E[S(\mathbf{x})] = 0$ and $\operatorname{cov}\{S(\mathbf{x}), S(\mathbf{x}')\} = \sigma^2 \rho(\mathbf{x} - \mathbf{x}')$;
(b) conditionally on S, the random variables Y_i, $i = 1, \ldots, n$, are mutually independent, with distributions $f_i\{y|S(\mathbf{x}_i)\} = f(y;\ M_i)$ specified by the values of the conditional expectations $M_i = E[Y_i|S(\mathbf{x}_i)]$;

(c) $h(M_i) = S(\mathbf{x}_i) + \mathbf{d}_i^{\mathrm{T}}\boldsymbol{\beta}$, for some known link function h, explanatory variables $\mathbf{d}_i = \mathbf{d}(\mathbf{x}_i)$ and parameters $\boldsymbol{\beta}$.

Under these assumptions, we call the resulting expression for the kriging predictor $\hat{S}(\mathbf{x}) = E[S(\mathbf{x})|\mathbf{Y}]$ the *generalized linear predictor* for $S(\mathbf{x})$. Because expectation is a linear operator, the form of the generalized linear predictor for any linear functional T is then given by equation (5).

We write $\mathbf{S} = (S(\mathbf{x}_1), \ldots, S(\mathbf{x}_n))^{\mathrm{T}}$ for the set of values of $S(\mathbf{x})$ at the sampling locations $\mathbf{x}_i$, and $\mathbf{S}^* = (S(\mathbf{x}_1^*), \ldots, S(\mathbf{x}_m^*))^{\mathrm{T}}$ for the corresponding set of values of $S(\mathbf{x})$ at the locations $\mathbf{x}_i^*$ for which predictions are required. Also, we let $g_k(\mathbf{s})$ denote the multivariate normal probability density function of the first k elements of $(\mathbf{S}, \mathbf{S}^*)$, and we let s_i denote the value of the ith element of this augmented vector, for each of $i = 1, \ldots, n+m$. Then, under assumptions (a)–(c) above, with all the parameters of the model taken to be fixed, the unconditional density of $\mathbf{Y}$ is given by the n-fold integral

$$f(\mathbf{y}) = \int \left\{ \prod_{i=1}^{n} f_i(y_i|s_i) \right\} g_n(\mathbf{s})\, \mathrm{d}\mathbf{s}, \tag{10}$$

and the unconditional density of $(\mathbf{S}^*, \mathbf{Y})$ by the n-fold integral

$$\int \left\{ \prod_{i=1}^{n} f_i(y_i|s_i) \right\} g_{n+m}(\mathbf{s})\, \mathrm{d}\mathbf{s}, \tag{11}$$

where $\mathrm{d}\mathbf{s} = \Pi_{i=1}^{n}\, \mathrm{d}s_i$. The conditional density of $(S_{n+1}, \ldots, S_{n+m})$ given $\mathbf{Y}$ is then the ratio of the two integrals in expressions (10) and (11). In particular, if we define the $(n+1)$-fold integral

$$E_{r,j} = \int s_{n+j}^{r} \left\{ \prod_{i=1}^{n} f_i(y_i|s_i) \right\} g_{n+m}(\mathbf{s})\, \mathrm{d}\mathbf{s}\, \mathrm{d}s_{n+j},$$

for positive integer r and $j = 1, \ldots, m$, then the generalized linear predictor and the prediction variance, for location $\mathbf{x}_j^*$, are given by

$$\hat{S}(\mathbf{x}_j^*) = E_{1,j}/f(\mathbf{y}) \tag{12}$$

and

$$V(\mathbf{x}_j^*) = E_{2,j}/f(\mathbf{y}) - \hat{S}(\mathbf{x}_j^*)^2.$$

Since n will be very large, it is clear that special methods will be needed for the approximate evaluation of $f(\mathbf{y})$, $E_{1,j}$ and $E_{2,j}$. By comparison note that disjunctive kriging only needs the evaluation of two-dimensional integrals.

In Section 4, we attack this problem by using MCMC methods (Smith and Roberts, 1993; Besag and Green, 1993; Gilks *et al.*, 1993), and adopting a Bayesian approach to inference. Since assumptions (a)–(c) specify explicit forms for the unconditional distribution of $\mathbf{S}$ and for the conditional distribution of $\mathbf{Y}$ given $\mathbf{S}$, we can use MCMC methods to simulate from the conditional distribution of $\mathbf{S}$ given $\mathbf{Y}$, and hence to estimate any expectation (or other functional) associated with this conditional distribution.

4. Inference and prediction

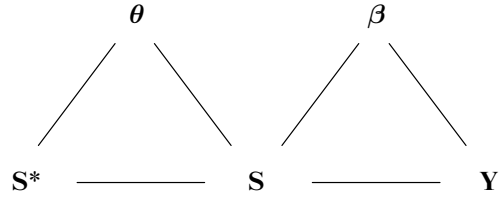
MCMC is a generic term to describe a collection of methods for simulating from complex multivariate distributions, and in particular from distributions whose probability density

function is analytically intractable. The basic idea is to set up a Markov chain whose transition probabilities *are* analytically tractable and which has the required multivariate distribution as its equilibrium distribution. By producing a sufficiently long run of this chain, we can therefore simulate a realization from the required distribution. By continuing to simulate from the chain after equilibrium has been attained, we can then obtain an arbitrarily large, albeit dependent, sample from the target distribution.

This idea has proved to be especially useful when the joint distribution of the observed data is specified only indirectly, via the specification of a set of lower dimensional marginal and conditional distributions. This makes MCMC methods a natural tool for Bayesian inference, and especially for models such as ours, where the marginal distribution of S and the conditional distribution of $\mathbf{Y}$ given S are specified explicitly. The Bayesian framework also provides a convenient way of incorporating parameter uncertainty into predictive inferences for the process S. We adopt independent proper uniform priors throughout, a consequence being that the resulting joint posterior distribution is proportional to the likelihood surface over the specified region.

Let $\boldsymbol{\theta}$ denote the set of parameters comprising the signal variance σ^2 and any further parameters in the specification of the correlation structure of S, and let $\boldsymbol{\beta}$ consist of all the regression parameters. Our objective is to use MCMC methods to estimate the model parameters, $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$, and to generate samples from the conditional distribution of $(\mathbf{S}, \mathbf{S}^*)$ given $\mathbf{Y}$ under assumptions (a)–(c) in Section 3. This allows us to obtain predictions for any functional of interest associated with this conditional distribution, while making proper allowance for uncertainty in the parameter estimates. Under a standard MCMC scheme we need to generate random samples from the posterior distribution of $(\boldsymbol{\theta}, \mathbf{S}, \boldsymbol{\beta})|\mathbf{Y}$ for inference and from the posterior distribution of $\mathbf{S}^*|(\mathbf{Y}, \boldsymbol{\theta}, \mathbf{S}, \boldsymbol{\beta})$ for prediction.

The implementation of our MCMC scheme requires sampling from the conditional distributions $\pi(\boldsymbol{\theta}|\mathbf{Y}, \mathbf{S}, \boldsymbol{\beta})$, $\pi(\boldsymbol{\beta}|\mathbf{Y}, \mathbf{S}, \boldsymbol{\theta})$ and $\pi(S_i|\mathbf{S}_{-i}, \mathbf{Y}, \boldsymbol{\theta}, \boldsymbol{\beta})$, where $\mathbf{S}_{-i}$ denotes the vector $\mathbf{S}$ with its ith element, $S_i = S(\mathbf{x}_i)$, removed. A schematic representation of the dependence structure of our model is as follows:



(13)

When the objective is inference about model parameters rather than prediction of $\mathbf{S}^*$, we can simplify this structure by dropping the $\mathbf{S}^*$-node. It is clear from the diagram that $\boldsymbol{\theta}$ is conditionally independent of $\mathbf{Y}$ given $\mathbf{S}$, and that $\boldsymbol{\beta}$ is conditionally independent of $\boldsymbol{\theta}$ given $\mathbf{S}$. So, for inference, a single cycle of the MCMC algorithm involves first sampling $\boldsymbol{\theta}|\mathbf{S}$, then $\mathbf{S}|(\mathbf{Y}, \boldsymbol{\theta}, \boldsymbol{\beta})$ and finally $\boldsymbol{\beta}|(\mathbf{Y}, \mathbf{S})$. Following assumption (b) in Section 3, we note that

$$p(\mathbf{Y}|\mathbf{S}, \boldsymbol{\beta}) = \prod_{j=1}^{n} f(y_j|s_j, \boldsymbol{\beta}), \tag{14}$$

where $f(y_j|s_j, \boldsymbol{\beta}) \equiv f(y; M_j)$. Using equation (14), the conditionals can now be written as

$$\pi(\boldsymbol{\theta}|\mathbf{Y}, \mathbf{S}, \boldsymbol{\beta}) = \pi(\boldsymbol{\theta}|\mathbf{S}) \propto p(\mathbf{S}|\boldsymbol{\theta})\, p(\boldsymbol{\theta}), \tag{15}$$

$$\pi(S_i|\mathbf{S}_{-i}, \mathbf{Y}, \boldsymbol{\theta}, \boldsymbol{\beta}) \propto p(\mathbf{Y}|\mathbf{S}, \boldsymbol{\beta})\, p(S_i|\mathbf{S}_{-i}, \boldsymbol{\theta})$$
$$= \left\{\prod_{j=1}^{n} f(y_j|s_j, \boldsymbol{\beta})\right\} p(S_i|\mathbf{S}_{-i}, \boldsymbol{\theta}), \tag{16}$$

$$\pi(\boldsymbol{\beta}|\mathbf{Y}, \mathbf{S}, \boldsymbol{\theta}) = \pi(\boldsymbol{\beta}|\mathbf{Y}, \mathbf{S}) \propto p(\mathbf{Y}|\mathbf{S}, \boldsymbol{\beta})\, p(\boldsymbol{\beta})$$
$$= \left\{\prod_{j=1}^{n} f(y_j|s_j, \boldsymbol{\beta})\right\} p(\boldsymbol{\beta}). \tag{17}$$

It follows from assumption (a) in Section 3 that $p(\mathbf{S}\,|\,\boldsymbol{\theta})$ in equation (15) has a multivariate normal density. As a consequence, $p(S_i|\mathbf{S}_{-i}, \boldsymbol{\theta})$ in equation (16) has a univariate Gaussian distribution. In principle, we could choose any prior distributions $p(\boldsymbol{\theta})$ and $p(\boldsymbol{\beta})$ in equations (15) and (17) respectively. However, we take $p(\boldsymbol{\theta})$ and $p(\boldsymbol{\beta})$ to be constants as we have chosen to use proper uniform priors.

We use the Metropolis–Hastings (MH) algorithm (see Smith and Roberts (1993)) to sample from the above conditional distributions. The MH scheme is not only straightforward to implement, but it is also a general purpose method. This makes it suitable for handling any form of conditional distribution in assumption (b) of Section 3. If we assumed a log-concave density for this conditional distribution, we could use the adaptive rejection sampling algorithm (Gilks and Wild, 1992) to sample from the conditional distributions. However, the resulting advantage in terms of computational efficiency would still need to be set against the increase in programming complexity and the restriction to the allowed class of models.

Our MCMC algorithm consists of the following steps.

Step 0: initial values need to be chosen for $\boldsymbol{\theta}$, $\boldsymbol{\beta}$ and $\mathbf{S}$ to start the algorithm. Arbitrary starting values for $\boldsymbol{\theta}$ may be chosen in the range specified by the prior. However, on assuming $\mathbf{S} \equiv 0$, our model reduces to a standard generalized linear model and the corresponding estimate of $\boldsymbol{\beta}$ from this model may be used as an initial value $\boldsymbol{\beta}^{(0)}$. We then set the starting value $s_i^{(0)}$ for each S_i by equating M_i to y_i for each i in the link function defined in assumption (c) in Section 3, to give $s_i^{(0)} = h(y_i) - \mathbf{d}_i^{\mathrm{T}}\boldsymbol{\beta}^{(0)}$.

Step 1: update all the components of the parameter vector $\boldsymbol{\theta}$—

(a) choose a new state $\boldsymbol{\theta}'$ uniformly from the parameter space specified by the prior;
(b) accept $\boldsymbol{\theta}'$ with probability $\Delta(\boldsymbol{\theta}, \boldsymbol{\theta}') = \min\{p(\mathbf{S}|\boldsymbol{\theta}')/p(\mathbf{S}|\boldsymbol{\theta}), 1\}$.

Step 2: update the signals $\mathbf{S}$—

(a) choose a new value S_i' for the ith component of $\mathbf{S}$ from the transition probability function $q(S_i, S_i') = p(S_i'|\mathbf{S}_{-i}, \boldsymbol{\theta})$;
(b) accept S_i' with probability $\Delta(S_i, S_i') = \min\{f(y_i|s_i', \boldsymbol{\beta})/f(y_i|s_i, \boldsymbol{\beta}), 1\}$;
(c) repeat (a) and (b) for all $i = 1, \ldots, n$.

Step 3: update all the elements of the regression parameter $\boldsymbol{\beta}$—

(a) choose a new value $\boldsymbol{\beta}'$ from some appropriate density $q(\boldsymbol{\beta}, \boldsymbol{\beta}')$;
(b) accept $\boldsymbol{\beta}'$ with probability

$$\Delta(\boldsymbol{\beta}, \boldsymbol{\beta}') = \min\left\{\frac{\prod_{j=1}^{n} f(y_j|s_j, \boldsymbol{\beta}')\, q(\boldsymbol{\beta}', \boldsymbol{\beta})}{\prod_{j=1}^{n} f(y_j|s_j, \boldsymbol{\beta})\, q(\boldsymbol{\beta}, \boldsymbol{\beta}')}, 1\right\}.$$

The choice of the transition kernel $q(\boldsymbol{\beta}, \boldsymbol{\beta}')$ in step 3 is problem specific; for an example, see Section 6.1. In practice, we update $\boldsymbol{\theta}$, $\mathbf{S}$ and $\boldsymbol{\beta}$ one component at a time, using the MH method to generate samples from the corresponding univariate conditional distributions.

We iterate steps 1–3 until the chain is judged to have reached its equilibrium distribution, at which point we introduce the following step.

Step 4: draw a random sample from the multivariate Gaussian distribution of

$$\mathbf{S}^*|(\mathbf{Y}, \boldsymbol{\theta}, \mathbf{S}, \boldsymbol{\beta}),$$

where $(\boldsymbol{\theta}, \mathbf{S}, \boldsymbol{\beta})$ are the values generated in steps 1–3. This step reduces to direct simulation from the Gaussian distribution of $\mathbf{S}^*|(\mathbf{S}, \boldsymbol{\theta})$, since our model implies that $\mathbf{S}^*$ is conditionally independent of both $\mathbf{Y}$ and $\boldsymbol{\beta}$, given $\mathbf{S}$. Specifically, it follows from assumption (a) in Section 3 that

$$\mathbf{S}^*|(\mathbf{S}, \boldsymbol{\theta}) \sim \mathrm{MVN}(\Sigma_{12}^{\mathrm{T}}\Sigma_{11}^{-1}\mathbf{S}, \Sigma_{22} - \Sigma_{12}^{\mathrm{T}}\Sigma_{11}^{-1}\Sigma_{12}), \tag{18}$$

where

$$\begin{aligned}\Sigma_{11} &= \mathrm{var}(\mathbf{S}),\\ \Sigma_{12} &= \mathrm{cov}(\mathbf{S}, \mathbf{S}^*),\\ \Sigma_{22} &= \mathrm{var}(\mathbf{S}^*).\end{aligned}$$

We then cycle over steps 1–4 as many times as necessary to obtain the required number of realizations from the distributions of $(\boldsymbol{\theta}, \mathbf{S}, \boldsymbol{\beta})|\mathbf{Y}$ and of $\mathbf{S}^*|(\mathbf{S}, \boldsymbol{\theta})$. After convergence we sample every kth realization of the chain. Increasing the value of k reduces the serial correlation in the resulting output sample. Step 4 is then only necessary at every kth cycle.

5. The variogram of the generalized linear prediction model

Despite the similarities between our model for generalized linear prediction and the generalized linear model, the diagnostic checks for the latter (McCullagh and Nelder, 1989) are inappropriate here owing to the spatial nature of the problem (Handcock and Wallis, 1994). However, a diagnostic for the second-moment structure of the fitted model is valuable and this is given by a comparison of the fitted and empirical variograms. In this section we discuss the theoretical form of the variogram for the generalized linear prediction model and its estimation.

Following from equation (8) and assumptions (a)–(c) in Section 3, a standard conditional expectation argument gives

$$\begin{aligned}C(\mathbf{u}) &= \tfrac{1}{2}\,\mathrm{var}\{Y(\mathbf{x})\} + \tfrac{1}{2}\,\mathrm{var}\{Y(\mathbf{x}+\mathbf{u})\} - \mathrm{cov}\{Y(\mathbf{x}), Y(\mathbf{x}+\mathbf{u})\}\\ &= \tfrac{1}{2}E_S[\mathrm{var}_Y\{Y(\mathbf{x})|S\}] + \tfrac{1}{2}E_S[\mathrm{var}_Y\{Y(\mathbf{x}+\mathbf{u})|S\}] + \tfrac{1}{2}\,\mathrm{var}_S\{E_Y[Y(\mathbf{x})|S]\}\\ &\quad + \tfrac{1}{2}\,\mathrm{var}_S\{E_Y[Y(\mathbf{x}+\mathbf{u})|S]\} - \mathrm{cov}_S\{E_Y[Y(\mathbf{x})|S], E_Y[Y(\mathbf{x}+\mathbf{u})|S]\},\end{aligned} \tag{19}$$

where we have used the fact that $E_S[\mathrm{cov}_Y\{Y(\mathbf{x}), Y(\mathbf{x}+\mathbf{u})|S\}] = 0$. Writing $M(\mathbf{x}) = E_Y[Y(\mathbf{x})|S]$ and $\tau^2(\mathbf{x}) = \mathrm{var}_Y\{Y(\mathbf{x})|S\}$, equation (19) simplifies to

$$\begin{aligned}C(\mathbf{u}) = \tfrac{1}{2}[&E_S\{\tau^2(\mathbf{x}) + \tau^2(\mathbf{x}+\mathbf{u})\} + \mathrm{var}_S\{M(\mathbf{x})\} + \mathrm{var}_S\{M(\mathbf{x}+\mathbf{u})\}]\\ &- \mathrm{cov}_S\{M(\mathbf{x}), M(\mathbf{x}+\mathbf{u})\}.\end{aligned} \tag{20}$$

Further simplification is possible only when additional assumptions are made as we illustrate through the following examples. More generally, the terms in expression (20) can be approximated from the MCMC sample output.

5.1. Example 1

In the case of linear kriging, i.e. when the conditional distribution of $Y(\mathbf{x})$ given $S(\mathbf{x})$ is Gaussian with mean $M(\mathbf{x}) = \beta + S(\mathbf{x})$ and variance $\tau^2(\mathbf{x}) = \tau^2$ for all $\mathbf{x}$, then equation (20) reduces to

$$C(\mathbf{u}) = \tau^2 + \sigma^2\{1 - \rho(\mathbf{u})\}. \tag{21}$$

The quantity τ^2 is often referred to as the nugget effect.

5.2. Example 2

When the distribution of $Y(\mathbf{x})$ given $S(\mathbf{x})$ is Poisson with mean $M(\mathbf{x}) = \exp\{\beta + S(\mathbf{x})\}$ and $\tau^2(\mathbf{x}) = M(\mathbf{x})$, equation (20) yields

$$C(\mathbf{u}) = \exp\left(\beta + \frac{\sigma^2}{2}\right) + \exp(2\beta + \sigma^2)[\exp(\sigma^2) - \exp\{\sigma^2\,\rho(\mathbf{u})\}]. \tag{22}$$

When σ is small, equation (22) gives, to the first order of approximation,

$$C(\mathbf{u}) \approx \tau_*^2 + \sigma_*^2\{1 - \rho(\mathbf{u})\}$$

with $\tau_*^2 = \exp(\beta + \sigma^2/2)$ and $\sigma_*^2 = \sigma^2\tau_*^4$. This shows that equation (22) has essentially the same structure as the Gaussian variogram (21), except that the Poisson assumption imposes a functional link between the parameters τ_* and σ_*.

In practice the fitted variogram can be obtained by either of two routes: either by plugging in estimates of $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$ obtained from the MCMC algorithm into an expression derived from equation (20), or by treating $C(\mathbf{u})$ as a parameter functional of interest, in which case we obtain a sample from the posterior distribution for $C(\mathbf{u})$ by plugging in realized values of $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$ from the MCMC sample. A point estimate of $C(\mathbf{u})$, the fitted variogram $\hat{C}(\mathbf{u})$ and associated central credibility intervals can then be extracted from the posterior distribution.

6. Applications

In all three applications in this section we assume that S is a zero-mean stationary Gaussian process with variance σ^2 and isotropic correlation function

$$\rho(u) = \exp\{-(\alpha u)^\delta\}, \tag{23}$$

for some $\alpha > 0$, $\sigma > 0$ and $0 < \delta < 2$; thus $\boldsymbol{\theta} = (\sigma, \alpha, \delta)$. Although $\delta = 2$ is sometimes used in linear kriging, expression (23) generates a strictly positive definite correlation matrix for S only if $0 < \delta < 2$. At the boundaries $\delta = 0$ and $\delta = 2$, the resulting correlation matrix becomes positive semidefinite and hence singular. Another critical distinction between $\delta < 2$ and $\delta = 2$ is that S is mean square continuous and mean square differentiable respectively. This is of importance because in linear kriging the predicted surface $\hat{S}$ has the same degree of analytic smoothness as the assumed correlation function of S.

In each application the uniform priors that are used for σ^2 and α have 0 as the lower bound to their range. This has the potential problem that in theory 0 is an absorbing state of the

Markov chain for each parameter (Besag *et al.*, 1991). This was not found to be a problem in practice as it did not occur in any of our applications, presumably because the probability of reaching the absorbing state is effectively 0.

In the remainder of the paper, we assume that the explanatory variables $\mathbf{d}_i$ have been mean corrected over the data sites, i.e. $\Sigma_i \mathbf{d}_i = \mathbf{0}$, and that distance is scaled by the maximum distance over the region of interest so that $0 \leqslant u \leqslant 1$ in equation (23).

6.1. Simulated case-study

In this study the spatial dimension is reduced to 1, i.e. $t \in [0, 1]$. There are 150 equally spaced observations over the whole interval. For location t the data are generated from the following model: $Y_t|S(t)$ is Poisson with mean $M(t)$, where

$$\log\{M(t)\} = \beta_0 + \beta_1 d(t) + S(t). \tag{24}$$

In equation (24), $S(t)$ is a zero-mean Gaussian process, and the uncentred known explanatory variable $d(t)$ is given by

$$d(t) = 5t + 0.2Z_t,$$

with Z_t, $t = 1, \ldots, 150$, being independent and identically distributed standard normal random variables. In addition, conditionally on $\mathbf{S}$ the Y_t are independent. The true parameter values are $\boldsymbol{\theta} = (1, 20, 1.8)$ and $\boldsymbol{\beta} = (\beta_0, \beta_1) = (-1, 1)$. The generated data $\mathbf{Y}$ and underlying signal $\mathbf{S}$ are shown in Fig. 1. The parameter values were chosen to give a high proportion of 0s in the data sample while retaining some clear spatial and systematic structure, as is evident from Fig. 1. We assume that interest is in both the regression parameter β_1 and the realization of S.

When we applied the MCMC algorithm of Section 4 to these data we experienced very slow convergence, even with good starting values. It proved necessary to reparameterize the

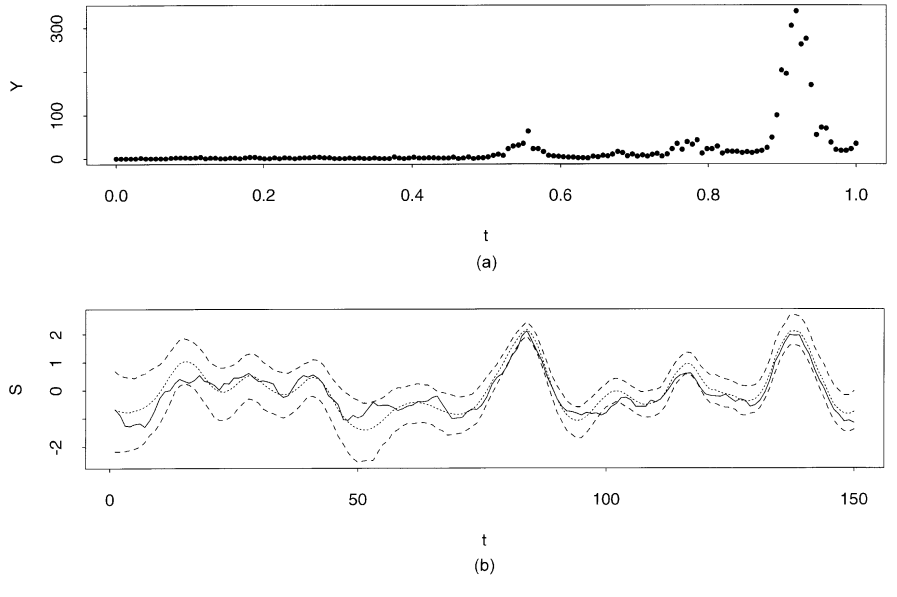


Fig. 1. (a) Observed data plotted against spatial location for the simulated data; (b) true signal (———) and its estimate (·····), along with the 95% central credibility interval (- - - -), given by generalized linear prediction for the simulated case-study

model slightly to obtain reasonable convergence. Specifically, when updating β_0 and β_1 we modified equation (24) to

$$\log\{M(t)\} = \beta_0^* + \beta_1 d(t) + S(t) - \bar{s},$$

where $\beta_0^* = \beta_0 + \bar{s}$, and $\bar{s}$ is the average of the current values of the predicted signals at the data sites. This reparameterization to (β_0^*, β_1) makes the parameters in the updating more orthogonal, which speeds up the algorithm. In step 3 of our algorithm, a Gaussian transition kernel is used for updating each of the components of $\boldsymbol{\beta}$. This choice speeds up the mixing of the chain. The sample mean and variance estimated from an initial run are used as parameters of the normal distribution from which the new value is drawn.

Fig. 2 shows MCMC samples of $(\boldsymbol{\theta}, \boldsymbol{\beta})$ from every 100th iteration when using independent uniform priors for $(\sigma^2, \alpha, \delta, \boldsymbol{\beta})$ over the range $[0, 10] \times [0, 50] \times [0.5, 1.95] \times [-2, 0] \times [0, 2]$.

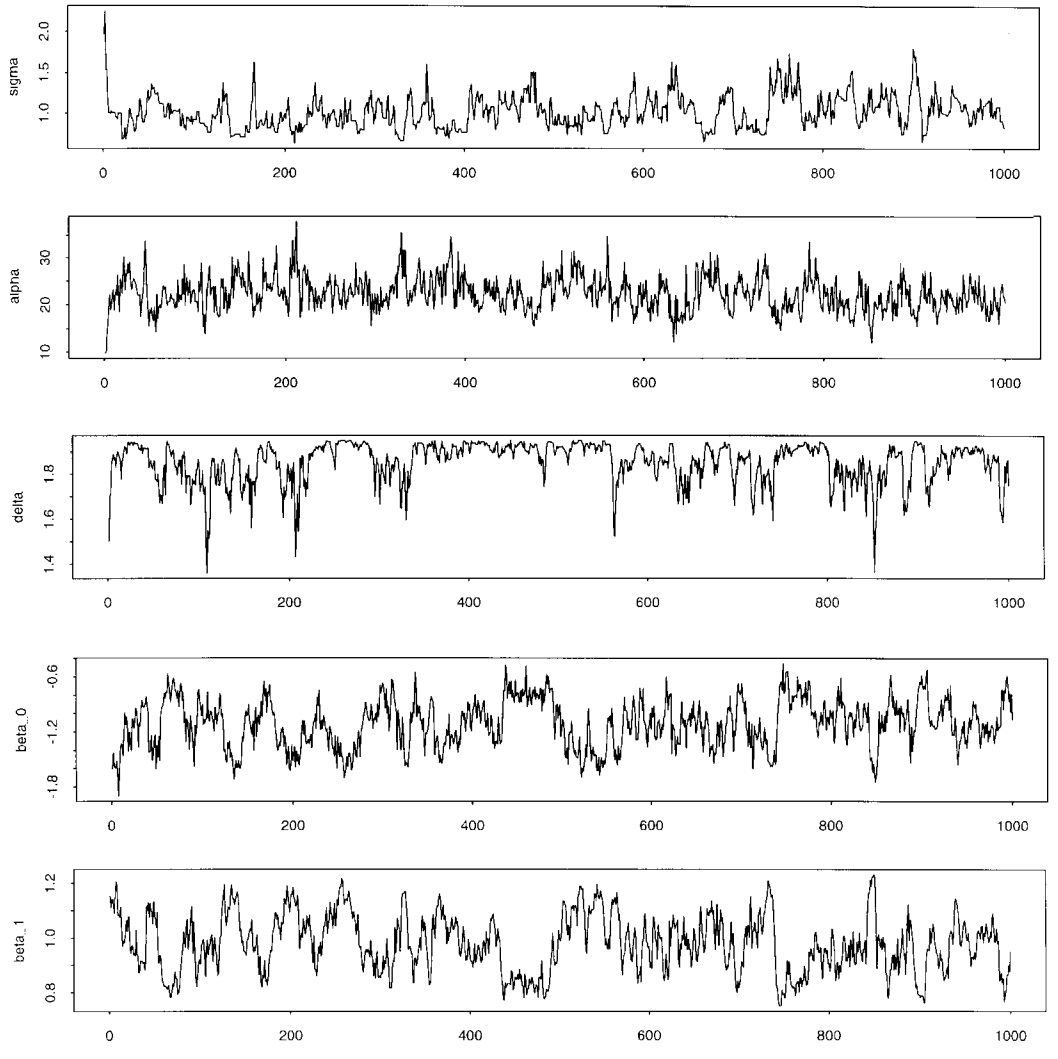


Fig. 2. Time series plots monitoring the MCMC output every 100th iteration for the simulated case-study

From Fig. 2 we ignore the first 300 samples, by which time convergence is judged to have occurred, and use all the subsequent samples to obtain the posterior distributions and various functionals of the parameters of interest. The marginal posterior for δ is quite asymmetric, with 7% of the MCMC sample within 0.01 of the upper bound on the range of the prior. This may suggest sensitivity to this prior specification; however, $1.95 < \delta < 2$ give only slightly smoother S-processes but have the drawback of numerical instability because var(**S**) is nearly singular, which leads to difficulty implementing steps 1 and 4 of our algorithm. For all the other $\boldsymbol{\theta}$- and $\boldsymbol{\beta}$-parameters, except σ, the marginal posteriors are reasonably symmetric, and all posterior distributions are consistent with the true values. For example, the marginal posterior density for β_1, shown in Fig. 3, has a mean which is close to 1. Similarly, the marginal posterior for β_0 has mean -1.08 and standard deviation 0.26. For each t, the marginal posterior distribution of $S(t)$ is also fairly symmetric, with Fig. 1(b) showing the posterior mean and the 95% central credibility interval for $S(t)$. The posterior mean appears to estimate the true signal reasonably well. In particular, it captures the size and positions of the main oscillations accurately, although it is generally smoother than the true signal.

To see how well the generalized linear prediction procedure performs in estimating the systematic component of the model we now consider estimation by using a standard generalized linear model, with link function as given in equation (24) but assuming incorrectly that $S(t) = 0$ for all t. The generalized linear model captures the systematic component of the generating model but ignores the stochastic mechanism which is responsible for the spatial dependence. Hence, inference about the regression parameters should give confidence intervals which are too narrow as the data are falsely assumed to be independent. This feature is seen clearly in Fig. 3, which shows the marginal posterior density for β_1, obtained by applying a slightly modified version of the MCMC algorithm of Section 4 to the generalized linear model. The true value is well outside any reasonable credibility interval. Similarly, β_0 has a marginal posterior density with mean at -1.56 and standard deviation 0.11, which are

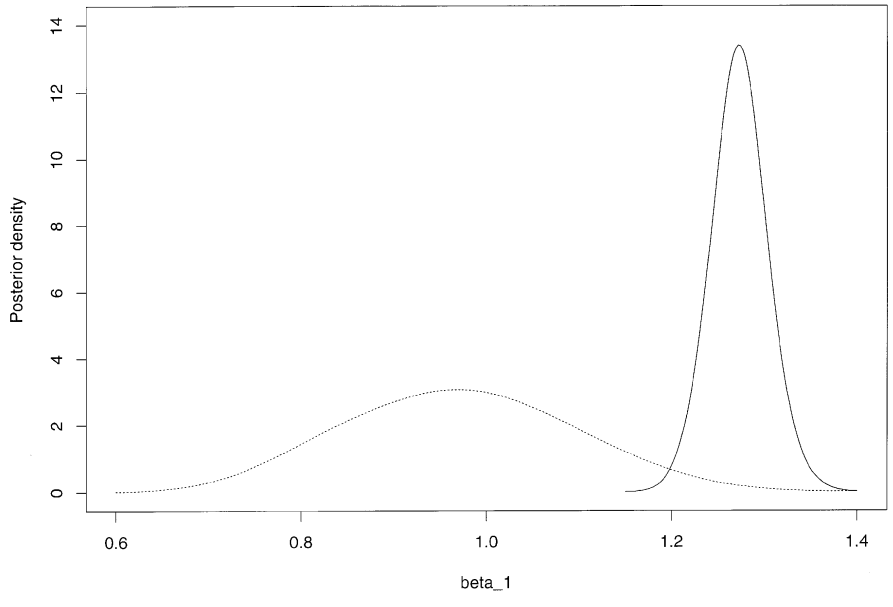


Fig. 3. Density plots showing the posterior density for β_1 obtained under the generalized linear model (———) and generalized linear prediction (·····) for the simulated case-study

inconsistent with the true value of -1. The explanation for this is that, within a single realization, there is partial confounding between the deterministic trend, $\beta_0 + \beta_1 t$, and the smooth stochastic variation about the trend, $S(t)$. This emphasizes that the $\boldsymbol{\beta}$-parameters must be interpreted conditionally on $S(t)$, rather than marginally. The need to distinguish between conditional and marginal regression parameters, which does not arise in linear Gaussian models, is well known in the context of generalized linear modelling for longitudinal data (see, for example, Diggle *et al.* (1994), chapter 7).

6.2. Radionuclide concentrations on Rongelap Island

Fig. 4 shows a map of Rongelap Island with the 157 sampling locations marked. These derive from an initial survey of measurements on a regular grid at a spacing of 200 m, which was later supplemented by measurements at a spacing of 40 m within four of the original grid squares to improve the ability to distinguish between large and small scales of spatial variation in the data. The empirical variogram for these data, under the assumption of isotropy, is shown in Fig. 5 as the broken curve. The behaviour of $\tilde{C}(u)$ near the origin suggests that the measurement error, which in our model-based approach represents Poisson sampling variation, accounts for a substantial proportion of the overall variation in the data.

Diggle *et al.* (1997) discussed the problems which arise in applying conventional geostatistical methodology to the Rongelap data. They converted the net counts at each sample location to counts per unit time, and then applied trans-Gaussian kriging with a log-transformation and correlation function $\rho(u)$ given by equation (23) with $\delta = 2$. They used the generalized least squares method of Cressie (1985) to estimate the covariance parameters, estimating α to be 8.9.

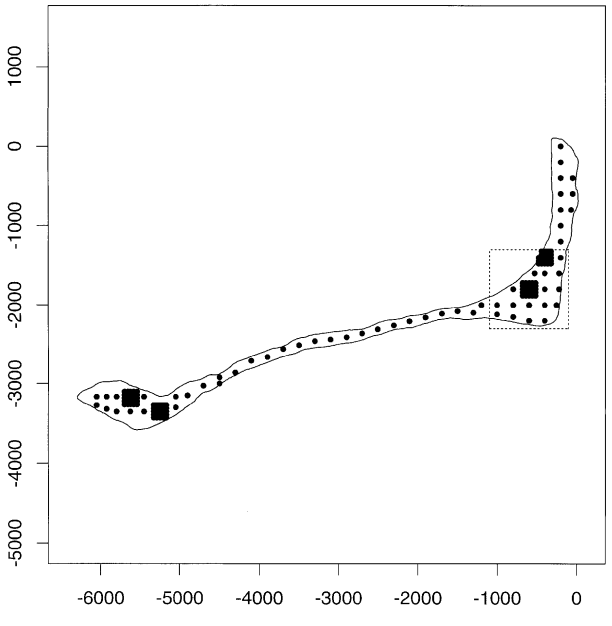


Fig. 4. Map of Rongelap Island with the 157 sampling locations marked: the distance scale is in metres relative to an arbitrary origin; the region indicated by the dotted box is shown in enlarged form in the inset plots in Figs 6 and 7

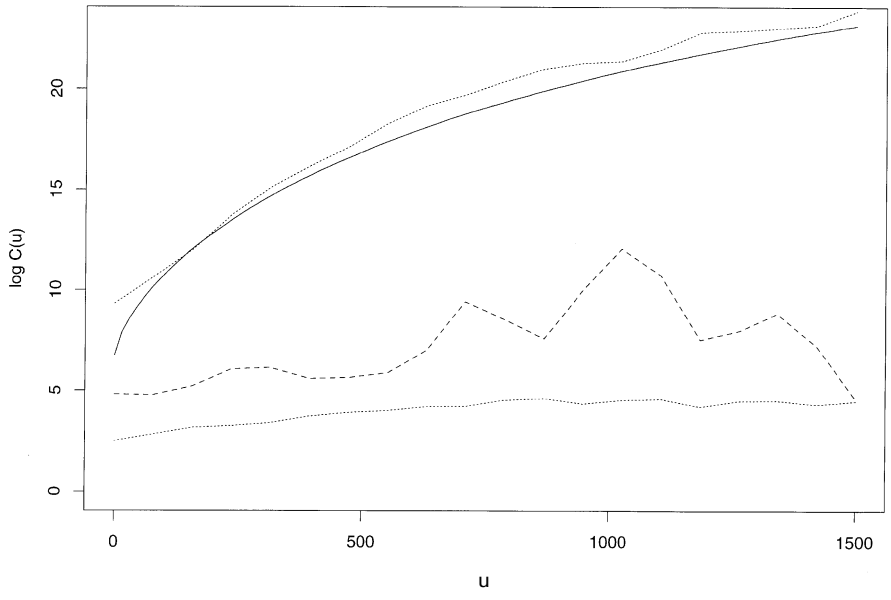


Fig. 5. Variogram plot on a log-scale (———, Poisson model; - - - -, empirical variogram; ·····, mean and 95% lower envelope constructed from 600 simulated data sets following the fitted Poisson model)

Fig. 6 reproduces their predicted surface $\exp\{\hat{\mu} + \hat{S}(\mathbf{x}) + V(\mathbf{x})/2\}$. Here, for data on the log-scale, $\hat{\mu}$ is the sample mean, $\hat{S}(\mathbf{x})$ is the linear kriging predictor (3) and $V(\mathbf{x})$ is the prediction variance (4). The map is derived from predictions made at the 157 sampling locations and a further 803 sites. This surface was judged by the originator of the data, Dr S. Simon, to be unsatisfactory because the differences between the original data and the smoothed predictions at the sample locations were implausibly large in relation to the known sampling errors of the field recording equipment. One possible explanation for this is that the assumed parametric model for the correlation structure of $S(\mathbf{x})$ is incorrect. Another is that the log-Gaussian kriging method makes inappropriate distributional assumptions about the data. We now present an analysis based on an alternative model which incorporates more plausible distributional assumptions and a more flexible correlation structure.

Our model is that conditionally on the realization of a stationary Gaussian process $S(\mathbf{x})$ the observed counts, Y_i, $i = 1, \ldots, n$, say, are mutually independent, Poisson-distributed random variables with means

$$M_i = t_i \exp\{\beta + S(\mathbf{x}_i)\},$$

where t_i denotes the duration of observation at location $\mathbf{x}_i$, β is a parameter and $S(\mathbf{x})$ is a zero-mean Gaussian process with correlation function of the form (23). We focus on the intensity of the radioactivity, and hence define

$$\lambda(\mathbf{x}) = \exp\{\beta + S(\mathbf{x})\}.$$

Since this example does not have any explanatory variables, we interpret β as a non-zero mean for the process S and update it in step 1 of the algorithm accordingly, i.e. we take $\boldsymbol{\theta} = (\beta, \sigma, \alpha, \delta)$. Starting values for $\boldsymbol{\theta}$ were based on the estimates obtained from the log-

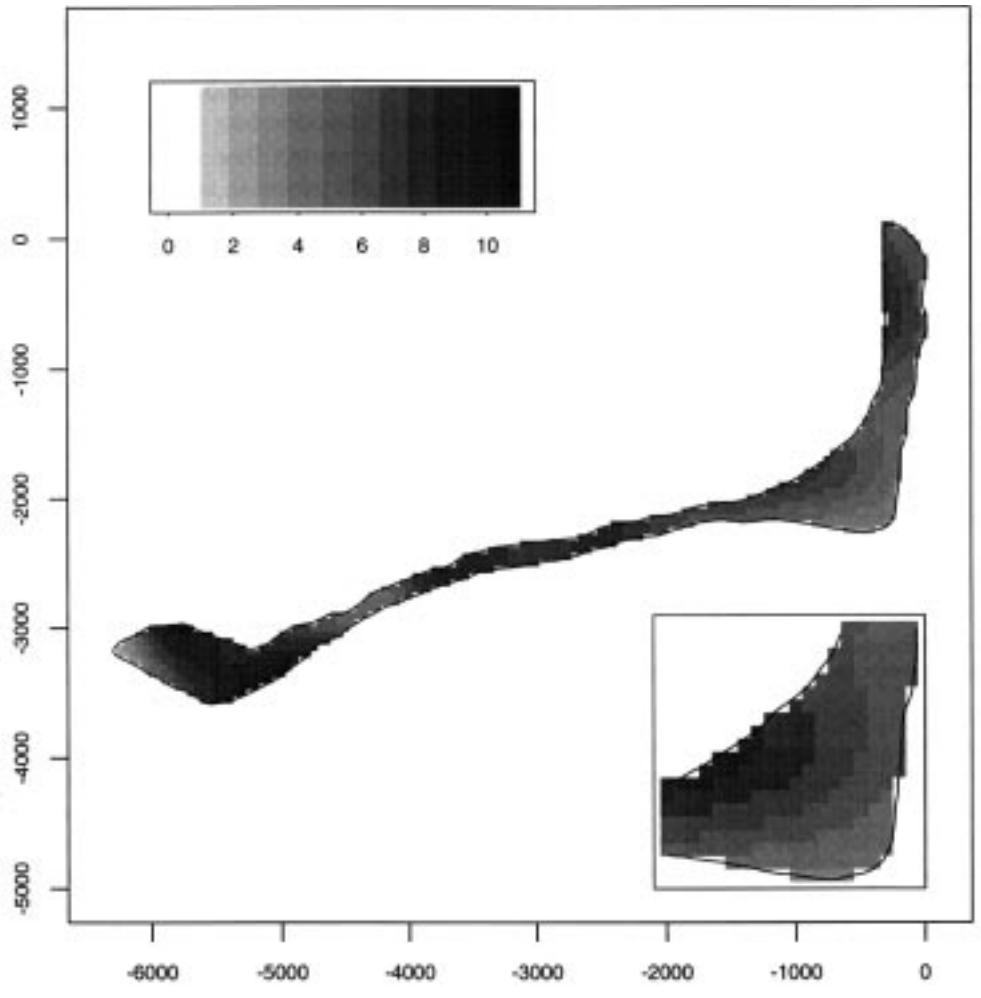


Fig. 6. Predicted intensity (i.e. counts per unit time) of ^{137}Cs over Rongelap Island by using log-Gaussian kriging: the distance scale is as in Fig. 4

Gaussian analysis in Diggle *et al*. (1997). We took independent uniform priors for $(\beta, \sigma^2, \alpha, \delta)$ over the range $[-3, 7] \times [0, 15] \times [0, 120] \times [0.1, 1.95]$. We ran the chain until convergence was judged to have occurred (approximately 1000 iterations), and then sampled the chain on every 100th of 50000 iterations to give a sample of 500 values of $\boldsymbol{\theta}$ together with $S(\mathbf{x})$ at each of the 960 locations used in Diggle *et al*. (1997). Our MCMC samples contain occasional instances of a very few consecutive ties for the α-parameter only, indicating reasonable acceptance rates for the MH steps. The marginal modes and means obtained for $\boldsymbol{\theta}$ are (1.7, 0.65, 4.7, 0.7) and (1.7, 0.89, 22.8, 0.7) respectively. These summary statistics illustrate a substantial skewness in the posterior for α.

We first assess the fit of the model through the variogram. By simulating independent replicated spatial samples from the fitted model, each with locations identical with those of the data, we can examine the variability of the empirical variogram and construct tolerance

intervals for it. Since $\tilde{C}(u)$, given by equation (9), is a consistent estimator, the sample mean for the replicated values of $\tilde{C}(u)$ will be close to $\hat{C}(u)$, and tolerance intervals should contain the fitted variogram. For diagnostic purposes we need to establish whether the data are consistent with the fitted variogram, i.e. whether the observed empirical variogram could have arisen from the fitted model. Fig. 5 shows the fitted variogram, together with the lower $2\frac{1}{2}\%$ tolerance level and the mean value from the simulations. The fit is adequate as judged by the 95% tolerance limits, but the width of these limits emphasizes the imprecision of $\tilde{C}(u)$ as an estimator for $C(u)$ when the data are relatively sparse, as in this example.

Our estimated values for α and δ are quite different from the estimates obtained by Diggle *et al.* (1997). However, α and δ are positively dependent so the effect of fixing $\delta = 2$ in Diggle *et al.* (1997) should be to increase α; this is consistent with the marginal mode but not the marginal mean of α. The change from $\delta > 1$ to $\delta < 1$ is most important as the estimates of the underlying process $S(\mathbf{x})$ no longer smooth the observed spatial variations in the data so heavily. For comparison with Fig. 6, in Fig. 7 we map $\lambda(\mathbf{x})$ using the sample mean of the $\lambda(\mathbf{x})$-values produced by the MCMC algorithm at each location. Although the basic structure is the same as in Fig. 6, the key difference is the substantially greater spatial variation over the island. The pattern and levels of the estimated surface $\lambda(\mathbf{x})$ now follow those of the observed data more closely. This reflects the knowledge of the investigator that the intensity at each location is very accurately measured, implying that the map should be in close agreement with the measured data at the sample locations.

The construction of Fig. 7 differs from that of Fig. 6 in three key ways: it is based on different distributional assumptions, it assumes a different correlation structure and it takes account of parameter uncertainty. We now examine the effect of the last of these differences. As discussed in Section 2, conventional geostatistical methods as applied by Diggle *et al.* (1997) ignore parameter uncertainty. Our procedure incorporates this uncertainty, and this affects both the estimated surface itself and, more noticeably, the prediction variances. As developed in Section 4, the algorithm naturally incorporates parameter uncertainty in the generalized linear prediction estimate. However, by simply repeating only step 4 once convergence is deemed to have occurred (with the parameters fixed at the marginal modes), we can investigate the effect of ignoring parameter uncertainty.

For the Rongelap data it is difficult to distinguish visually the maps of the posterior mean of $\lambda(\mathbf{x})$ produced with fixed and with varying parameters. However, potentially important differences do exist. At the data sites there is essentially no loss in taking the parameters as known since the data themselves are highly informative. However, for the prediction sites, at which data are not observed, estimates obtained with varying parameters are on average approximately 5% larger, rarely being smaller, and have prediction standard deviations which are 80% larger on average. Hence, ignoring parameter estimation does not change the predicted surface appreciably, i.e. the mean of $\lambda(\mathbf{x})|\mathbf{Y}$, but does change the variance of this conditional distribution substantially.

We complete the study of the Rongelap data by examining scientific questions about the radioactivity which illustrate the flexibility of our procedures. In each case the quantity of interest is a functional of $\lambda(\mathbf{x})$ and the distinction between parameters being fixed or varying will generally be important. We use varying parameters throughout, but in one case we also illustrate the effects of parameter uncertainty.

One critical issue concerns the value and location of the maximum level of radioactivity on the island. Thus, we are interested in $\max_{\mathbf{x}}\{\lambda(\mathbf{x})\}$ and in the spatial location of $\max_{\mathbf{x}}\{\lambda(\mathbf{x})\}$, $\mathbf{x}_{\max}$ say, in both cases conditionally on the data $\mathbf{Y}$. The power of the MCMC approach is evident here, as the distribution of $\mathbf{x}_{\max}$ can be estimated by the proportion of realizations of

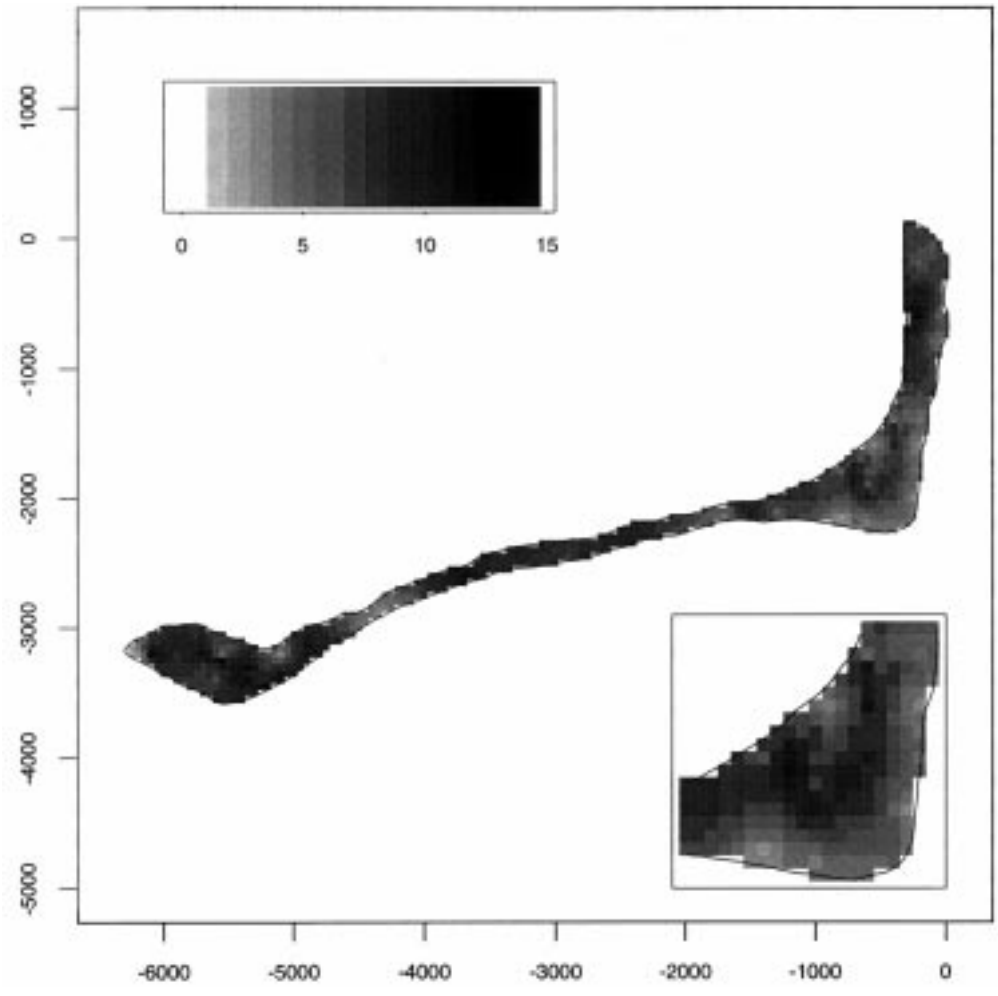


Fig. 7. Predicted intensity (i.e. counts per unit time) of ^{137}Cs over Rongelap Island by using generalized linear prediction: the distance scale is as in Fig. 4

the MCMC sample which give $\mathbf{x}_{\max}$ at a particular site. This is illustrated in Fig. 8, which shows the histogram of the frequency of the maximum, over the 500 realizations. Four distinct regions of the island, which are of specific concern, can be identified from Fig. 8. More critical is $\max_{\mathbf{x}}\{\lambda(\mathbf{x})\}$, for which the empirical distribution of the maximum from each realization of the MCMC spatial sample gives the posterior distribution estimate. This estimate is shown in Fig. 9, with and without taking account of uncertainty in the parameters $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$. Fig. 9 shows that ignoring parameter estimation has a major effect on the nominal precision of the estimation. The crude approach to finding the maximum, based on standard kriging methods, would be to take $\max_{\mathbf{x}}\{\hat{\lambda}(\mathbf{x})\}$. The corresponding values, derived from Figs 6 and 7, are 11 and 15 respectively; these values are in the lower tail of the posterior distribution of the maximum shown in Fig. 9(a).

In a general assessment of the habitability of the island, the proportion of the island for

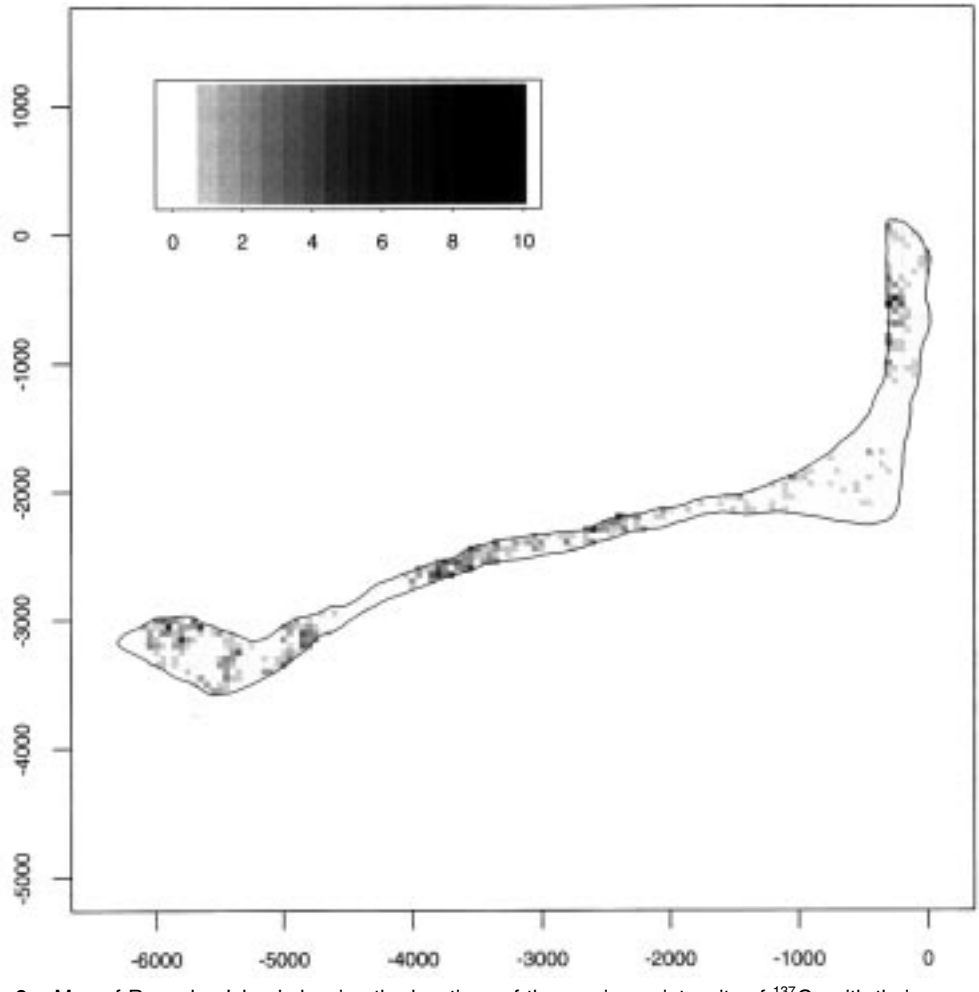


Fig. 8. Map of Rongelap Island showing the locations of the maximum intensity of ^{137}Cs with their respective frequencies of occurrence as a grey scale: the distance scale is as in Fig. 4

which the intensity of radioactivity exceeds a critical level, c say, is important. Fig. 9(b) shows the posterior distribution of this proportion in the form of a survivor function $p(c)$. Again, evaluation is simple using the MCMC method. The proportion of the predicted surface above c is obtained for every realization, and it is then straightforward to calculate the mean and the 95% central credibility interval.

Finally, suppose that the inhabitants agree to return to the island provided that they are reasonably confident that the intensity is below a critical level c in the regions of the island where they live. Then we require a central credibility set, at the 95% level say, for those regions of the island where the intensity is above the critical level, i.e. the set of $\mathbf{x}$ such that

$$\Pr\{\lambda(\mathbf{x}) > c|\mathbf{Y}\} > 0.05.$$

Lindgren and Rychlik (1994) considered a similar problem of obtaining confidence regions

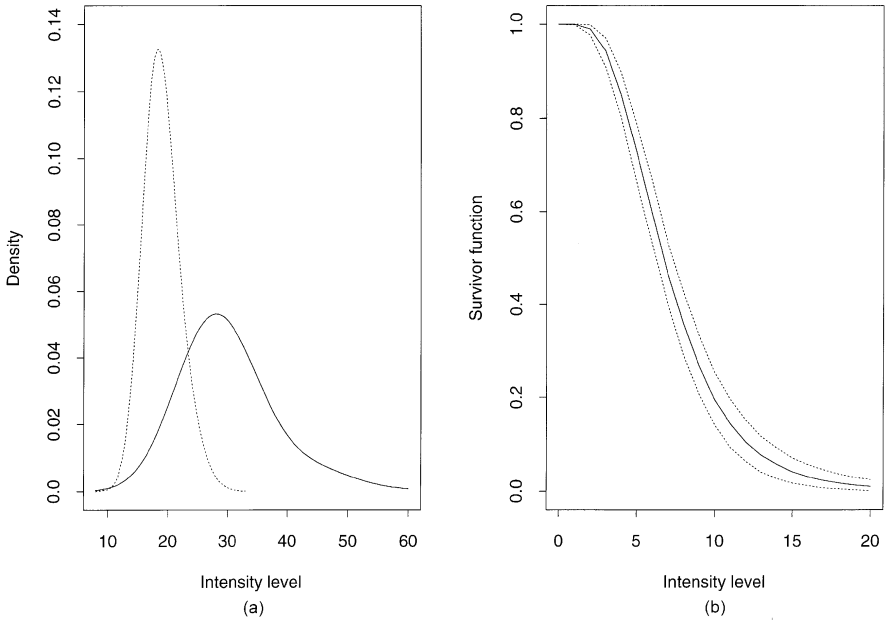


Fig. 9. (a) Kernel density estimate of the maximum intensity of ^{137}Cs (———, with varying parameters; ·····, with fixed values of the parameters); (b) proportion of the island above a given level of intensity with the corresponding pointwise 95% central credibility limits

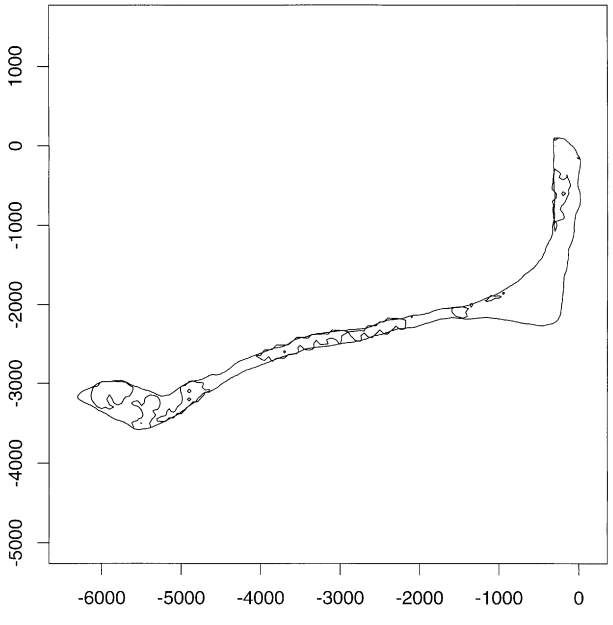


Fig. 10. Contour plot at the value 0.05 of the probability of ^{137}Cs intensity exceeding 15 counts per unit time over Rongelap Island: the distance scale is as in Fig. 4

for contour levels for model (2) with parameters fixed. However, despite the additional complexities in our model, the approach here is much more straightforward owing to the use of MCMC sampling. For each location on the predicted map, the proportion of realizations with intensity exceeding $c = 15$ is evaluated, and those sites at the 5% level are contoured to give Fig. 10. Not surprisingly, Fig. 10 and Fig. 8 are qualitatively similar, but Fig. 10 is more directly relevant to the problem of identifying regions of the island which have high intensity levels.

6.3. Campylobacter infections in north Lancashire and south Cumbria

The data for our third application consist of unit postcode locations and dates of onset of all recorded cases of campylobacter, salmonella and cryptosporidia between April 29th, 1991, and December 31st, 1994, in postcode sectors LA8, LA9, LA10, LA21, LA22 and LA23. Extrabinomial variation in these data could arise from two distinct sources: possible spatial variation in the relative risk of campylobacter infections and non-spatial variation because two or more people living at the same address may contract the disease from a common source of infection. This analysis focuses on the spatial variation in outbreaks, with the non-spatial variation accounted for by declustering the data before the spatial analysis. Exact declustering is not possible as individual addresses are not recorded for confidentiality. Approximate declustering therefore consisted of replacing any group of cases with the same unit postcode and dates of onset within 5 days of each other by a single case; the 5-day threshold was chosen on clinical rather than statistical grounds, based on the approximate latent period of the disease. The declustered data then contained 234 cases of campylobacter among 399 cases of enteric infections at 248 different unit postcode locations in the prescribed region as shown in Fig. 11.

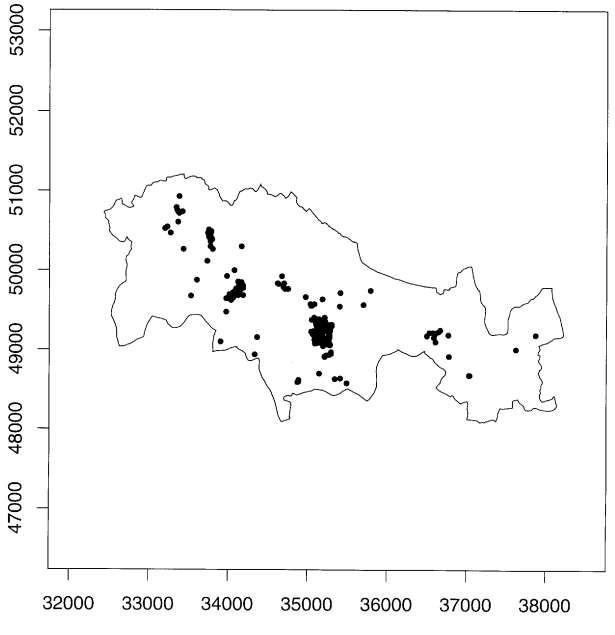


Fig. 11. Map of postcode sectors LA8, LA9, LA10, LA21, LA22 and LA23 showing locations of 248 unit postcodes at which cases of enteric infection were recorded: the grid values correspond to the national grid references, with each unit corresponding to 10 m

We model the declustered data as a set of conditionally independent Bernoulli trials with spatially varying 'success' probability

$$\log\left\{\frac{P(\mathbf{x})}{1-P(\mathbf{x})}\right\} = \beta + S(\mathbf{x}) \tag{25}$$

where $S(\mathbf{x})$ is a zero-mean stationary Gaussian process with correlation function given by equation (23). Equivalently, if Y_i, $i = 1, \ldots, n$, are the numbers of campylobacter cases at the $n = 248$ locations and there are m_i cases of enteric infection at location i, then Y_i is a binomially distributed random variable with parameters m_i and $p_i = P(\mathbf{x}_i)$, where $P(\mathbf{x}_i)$ is specified by equation (25).

Before attempting a spatial analysis, we need to confirm that the declustered data still exhibit significant extrabinomial variation. The residual deviance from a binomial model

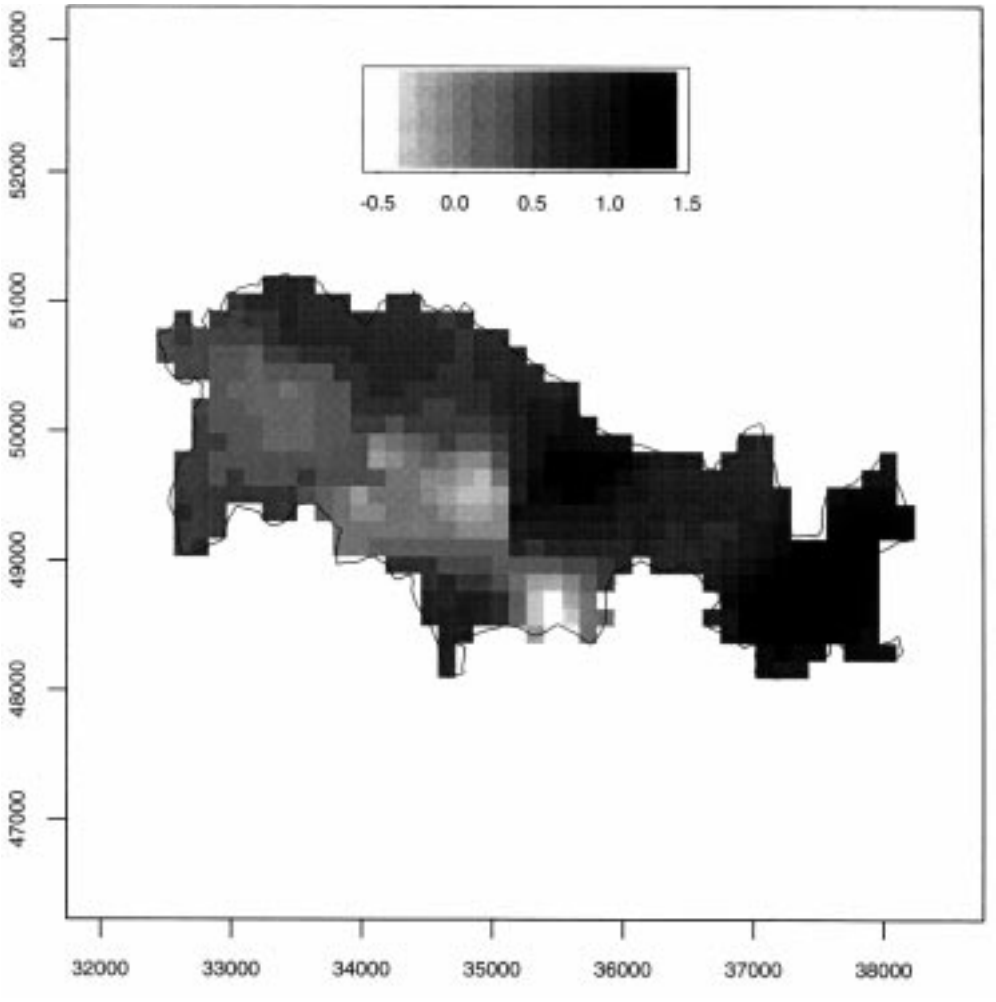


Fig. 12. Map of the estimated spatial variation in the log-odds that a case of enteric infection is due to campylobacter, relative to the average log-odds for the study region as a whole: the distance scale is as in Fig. 11

with constant $S(\mathbf{x})$ is 361.4 on 247 degrees of freedom, which is overwhelmingly significant ($p \approx 3 \times 10^{-6}$).

In implementing the MCMC algorithm for this problem, we experienced convergence problems when all parameters were allowed to vary, apparently because of a near non-identifiability involving δ and α. We therefore fixed $\delta = 1$ in equation (23) and allowed β, σ and α to vary. As in Section 6.2, we took β to be the expectation of $S(\mathbf{x})$ and updated it as an element of $\boldsymbol{\theta} = (\beta, \sigma, \alpha)$. Here we used independent uniform priors for $(\beta, \sigma^2, \alpha)$ over the range $[-1.5, 3.5] \times [0, 7] \times [0, 50]$. From the simulated chain we obtained 500 samples of $\boldsymbol{\theta}$, with each sample taken every 20th iteration after the time at which we judged that the chain had converged. The resulting marginal modes and means of the posterior were $\boldsymbol{\theta} = (0.62, 1.0, 6.5)$ and $(0.68, 1.2, 12.6)$ respectively. We then used the data to obtain predictions $\hat{S}(\mathbf{x})$ computed as sample means on a fine grid of 509 sites covering the whole region of interest. The resulting predictions are shown in Fig. 12. The observed variation in $\hat{S}(\mathbf{x})$ of $(-0.46, 1.42)$ corresponds to substantial spatial variation in the proportion of campylobacter cases, from 0.39 to 0.81. The map of $\hat{S}(\mathbf{x})$ is to be interpreted as an estimate of the spatial variation in the log-odds that a case of enteric infection is due to campylobacter, measured relative to the overall average log-odds. The areas of highest estimated log-odds are all rural in character, which is consistent with the suggestion that farms may provide an environmental reservoir of campylobacter organisms (Jones and Telford, 1991), although it must be acknowledged that the data analysed here are somewhat sparse.

In this example, we chose to account for non-spatial extrabinomial variation by a declustering method which, although somewhat *ad hoc* from a statistical point of view, was related to the particular biomedical context in which the data arose. An alternative approach would have been to add a non-spatial component to the model for $S(\mathbf{x})$. This would modify equation (25) to

$$\log\left\{\frac{P(\mathbf{x})}{1 - P(\mathbf{x})}\right\} = \beta + S(\mathbf{x}) + Z \tag{26}$$

where $Z \sim N(0, \tau^2)$ is independent of $S(\mathbf{x})$ and is realized independently at each unit postcode location. The Z-term in equation (26) is the nugget effect in classical geostatistics. In disease mapping, its inclusion corresponds to using a 'convolution Gaussian prior' (Mollié, 1996).

7. Discussion

As noted earlier, the methodology described in this paper has close connections with generalized linear mixed modelling (Breslow and Clayton, 1993) and Bayesian image restoration (Besag *et al.*, 1991). We have chosen to model spatial variation by using stationary Gaussian processes with a continuous index set. This follows the standard practice of classical geostatistics and allows us to make predictions at arbitrary locations within the study region without redefining our underlying model. When the spatial index set can be represented by a finite number of locations, Markov random field models would be computationally more convenient as they are defined through an explicit representation of the local dependence structure among the discrete set of locations considered. Arguably, this can always be done in practice by discretizing a continuous spatial region, although in general there would then be no easy way to translate a parametric model defined at one level of spatial resolution into an equivalent model at a finer resolution.

Neither the Gaussian assumption for the process S nor the generalized linear assumption for the distribution of $\mathbf{Y}$ conditional on S is essential for an MCMC implementation, but this

framework appears to be sufficiently general to accommodate a wide range of practical applications without being entirely nebulous. We favour abandoning either or both of these assumptions in applications only when there is a physical motivation for other specific assumptions to replace them. An important class of models which assumption (b) of Section 3 excludes is when the distribution of $Y_i|S(\mathbf{x}_i)$ depends on parameters ϕ as well as the mean $M(\mathbf{x}_i)$, a special case being model (2) with $\phi = \tau$. However, inference for such an extended formulation of the model is easily incorporated within the MCMC structure of Section 4 by treating ϕ as a component of the β-parameter; see Papritz and Moyeed (1997).

Similarly, although we have given prominence to the prediction of $S(\mathbf{x})$ to provide a direct extension of conventional geostatistics, the ability to make predictions for arbitrary functionals of S is an important advantage of the MCMC approach.

The use of MCMC methods is critical to the application of these methods. Our experience has been that the most important stage in implementing these is in the choice of parameterization of the model and the updating steps. In the development of our algorithm, we often found apparent convergence of the routine to a false equilibrium. From our experience we would strongly advise the users of MCMC methods to calibrate and test routines carefully before applying them. Throughout we have assessed convergence of the MCMC algorithm from visual and time series analysis of the sample output and by examining the sensitivity to a range of starting values and convergence periods. More formal methods are increasingly being adopted (see Cowles and Carlin (1996)), although these act as additional diagnostics, rather than as alternatives, to those used here.

A major benefit of the generalized linear prediction model proposed in this paper is the distributional extension of the standard trans-Gaussian kriging model. One consequence of this is that a functional link is imposed between the parameters for the spatial variation and measurement error, whereas for existing models, such as equation (2), these are unrelated parameters. If the assumed model is correct this can significantly improve inference and prediction, as appears to be the case for the Rongelap analysis. Otherwise, it can be problematic, as is seen by the following heuristic argument. Suppose that the underlying Gaussian process is $S(\mathbf{x}) = S_1(\mathbf{x}) + S_2(\mathbf{x})$, where S_1 and S_2 are independent zero-mean Gaussian processes exhibiting short and longer range spatial dependence respectively. In other words, the covariance structure of the assumed model is incorrect. Estimation under the standard trans-kriging model will typically absorb the spatial variation of S_1 into the measurement error parameter τ^2 and so to a first approximation does not bias the estimation of the spatial variation of S at longer distances, nor the kriging predictor. Under our generalized linear prediction formulation, variation of S_1 influences the estimation of the measurement error distribution, and so could either bias the mean effect model or the spatial variation through a trade-off between the linked parameters. As with all model-based approaches to inference, generalized linear prediction is a two-edged sword, requiring its users to address the assumptions made more critically than in the case of nonparametric smoothing methods. In our applications, the data show no evidence for additional short-range dependence, although this may be due to the limited information that is available in the data to identify such characteristics.

Acknowledgements

We thank Dr S. Simon of the Marshall Islands National Radiological Survey for providing the Rongelap data, Dr D. Telford of Lancaster Royal Infirmary for the campylobacter data and the Engineering and Physical Sciences Research Council for financial support.

References

Armstrong, M. and Matheron, G. (1986a) Disjunctive kriging revisited: part I. *Math. Geol.*, **18**, 711–728.

———(1986b) Disjunctive kriging revisited: part II. *Math. Geol.*, **18**, 729–742.

Besag, J. and Green, P. J. (1993) Spatial statistics and Bayesian computation. *J. R. Statist. Soc.* B, **55**, 25–37.

Besag, J., York, J. and Mollié, A. (1991) Bayesian image restoration, with two applications in spatial statistics (with discussion). *Ann. Inst. Statist. Math.*, **43**, 1–59.

Breslow, N. E. and Clayton, D. G. (1993) Approximate inference in generalized linear mixed models. *J. Am. Statist. Ass.*, **88**, 9–25.

Clayton, D. G. and Kaldor, J. (1987) Empirical Bayes estimates of age-standardized relative risks for use in disease mapping. *Biometrics*, **43**, 671–681.

Cowles, M. K. and Carlin, B. P. (1996) Markov chain Monte Carlo convergence diagnostics: a comparative review. *J. Am. Statist. Ass.*, **91**, 883–904.

Cressie, N. (1985) Fitting variogram models by weighted least squares. *J. Int. Ass. Math. Geol.*, **17**, 563–586.

———(1991) *Statistics for Spatial Data*. New York: Wiley.

———(1994) Comment on the paper by Handcock and Wallis. *J. Am. Statist. Ass.*, **89**, 379–382.

Diggle, P. J., Harper, L. and Simon, S. (1997) A geostatistical analysis of residual contamination from nuclear weapons testing. In *Statistics for the Environment 3* (eds V. Barnett and K. F. Turkman), pp. 89–107. Chichester: Wiley.

Diggle, P. J., Liang, K.-Y. and Zeger, S. L. (1994) *The Analysis of Longitudinal Data*. Oxford: Clarendon.

Freulon, X. (1994) Conditional simulation of a Gaussian random vector with non linear and/or noisy observations. In *Geostatistical Simulations* (eds M. Armstrong and P. A. Dowd), pp. 57–71. Dordrecht: Kluwer.

Gilks, W. R., Clayton, D. G., Spiegelhalter, D. J., Best, N. G., McNeil, A. J., Sharples, L. D. and Kirby, A. J. (1993) Modelling complexity: applications of Gibbs sampling in medicine. *J. R. Statist. Soc.* B, **55**, 39–52.

Gilks, W. R. and Wild, P. (1992) Adaptive rejection sampling for Gibbs sampling. *Appl. Statist.*, **41**, 337–348.

Handcock, M. S. and Stein, M. L. (1993) A Bayesian analysis of kriging. *Technometrics*, **35**, 403–410.

Handcock, M. S. and Wallis, J. R. (1994) An approach to statistical spatial-temporal modelling of meteorological fields. *J. Am. Statist. Ass.*, **89**, 368–390.

Isaaks, E. H. and Srivastava, R. M. (1989) *An Introduction to Applied Geostatistics*. Oxford: Oxford University Press.

Jones, K. and Telford, D. (1991) On the trail of a seasonal microbe. *New Scient.*, Apr. 6th, 36–39.

Journel, A. G. (1983) Nonparametric estimation of spatial distributions. *J. Int. Ass. Math. Geol.*, **9**, 563–586.

Journel, A. G. and Huijbregts, C. J. (1978) *Mining Geostatistics*. London: Academic Press.

Laslett, G. M. (1994) Kriging and splines: an empirical comparison of their predictive performance in some applications (with discussion). *J. Am. Statist. Ass.*, **89**, 391–409.

Lawson, A. B., Biggeri, A. and Lagazio, C. (1996) Modelling heterogeneity in discrete spatial data models via MAP and MCMC methods. In *Proc. 11th Int. Wrkshp Statistical Modelling* (eds A. Forcina, G. M. Marchetti, R. Hatzinger and G. Galmacci), pp. 240–250. Citta di Castello: Graphos.

Le, N. D. and Zidek, J. V. (1992) Interpolation with uncertain spatial covariances: a Bayesian alternative to kriging. *J. Multiv. Anal.*, **43**, 351–374.

Lindgren, G. and Rychlik, I. (1994) How reliable are contour curves — confidence sets for level contours. *Technical Report 3102*. University of Lund, Lund.

Mardia, K. V., Kent, J. T. and Bibby, J. M. (1979) *Multivariate Analysis*. London: Academic Press.

Mardia, K. V. and Marshall, R. J. (1984) Maximum likelihood estimation of models for residual covariance in spatial regression. *Biometrika*, **72**, 135–146.

Mardia, K. V. and Watkins, A. J. (1989) On multimodality of the likelihood in the spatial linear model. *Biometrika*, **76**, 289–295.

Matérn, B. (1960) Spatial variation. *Report*. Statens Skogsforsningsinstitut, Stockholm.

Matheron, G. (1970) The theory of regionalized variables and its applications. *Cah. Centr. Morph. Math. Fontainebleau*, no. 5.

———(1976) A simple substitute for conditional expectation: the disjunctive kriging. In *Advanced Geostatistics in the Mining Industry* (eds M. Guarascio, M. David and C. Huijbregts), pp. 221–236. Dordrecht: Reidel.

McCullagh, P. and Nelder, J. A. (1989) *Generalized Linear Models*, 2nd edn. London: Chapman and Hall.

Mollié, A. (1996) Bayesian mapping of disease. In *Markov Chain Monte Carlo in Practice* (eds W. R. Gilks, S. Richardson and D. J. Spiegelhalter), pp. 359–379. London: Chapman and Hall.

Nelder, J. A. and Wedderburn, R. W. M. (1972) Generalized linear models. *J. R. Statist. Soc.* A, **135**, 370–384.

Papritz, A. and Moyeed, R. A. (1997) Empirical validation of linear and nonlinear methods for spatial point prediction. *Technical Report*. Institute of Terrestrial Ecology, Eidgenössiche Technische Hochschule, Zurich.

Raftery, A. E. and Banfield, J. D. (1991) Stopping the Gibbs sampler, the use of morphology, and other issues in spatial statistics. *Ann. Inst. Statist. Math.*, **43**, 32–43.

Ripley, B. D. (1981) *Spatial Statistics*. New York: Wiley.

Smith, A. F. M. and Roberts, G. O. (1993) Bayesian computation via the Gibbs sampler and related Markov chain Monte Carlo methods. *J. R. Statist. Soc.* B, **55**, 3–23.

Vecchia, A. V. (1988) Estimation and model identification for continuous spatial processes. *J. R. Statist. Soc.* B, **50**, 297–312.

———(1992) A new method of prediction for spatial regression models with correlated errors. *J. R. Statist. Soc.* B, **54**, 813–830.
Warnes, J. J. and Ripley, B. D. (1987) Problems with likelihood estimation of covariance functions of spatial Gaussian processes. *Biometrika*, **74**, 640–642.
Whittle, P. (1963) *Prediction and Regulation*. London: English Universities Press.
Wolpert, R. L. and Ickstadt, K. (1997) Poisson/gamma random field models for spatial statistics. To be published.
Zimmerman, D. (1989) Computationally efficient restricted maximum likelihood estimation of generalized covariance functions. *Math. Geol.*, **21**, 655–672.

Discussion on the paper by Diggle *et al.*

R. Webster (*Rothamsted Experimental Station, Harpenden*)

The origins of geostatistics can be identified in Kolmogorov's pioneering search for a method of optimal interpolation in the 1930s. His solution (Kolmogorov, 1941) we can recognize as ordinary kriging, though it was impracticable at the time for want of computers. The advent of the last and the economic advantages to be gained in mining and petroleum engineering propelled geostatistics forwards in the 1960s. As the authors state, the development of the subject is due largely to G. Matheron and his associates at the Paris School of Mines and in particular his seminal thesis (Matheron, 1965), though let us not forget the contribution of D. G. Krige himself, e.g. Krige (1966). Now geostatistics comprises a large body of theory, technique and case history in ever widening fields, including substantial application in my own of soil science. It is by no means restricted to linear methods; it includes important non-linear relationships as in grade–cut-off curves for evaluating profitability in mining (see Rivoirard (1994)) and the costs of environmental pollution and remediation (Goovaerts *et al.*, 1997).

I was especially interested in the authors' application of their technique to the infection by campylobacter. On the face of it this seems to be a very different kind of problem from those tackled by geostatisticians, even a non-problem. After all, you have complete information: you know the number of people living in each postal zone, and you know how many cases there were of the disease in any given spell and hence the rate of infection. It is not immediately obvious what more you could extract from the data. What makes the data interesting is that they can be used to estimate the risk of infection, and this is the line that the authors pursue.

We faced a similar situation a few years ago with another disease, namely cancer among children in the West Midlands of England. In some ways this is a little more straightforward because as far as we know cancer is not passed from person to person, and so the only source of variation in addition to the binomial distribution is spatial, presumably from as yet unidentified environmental factors. We could therefore model the risk of a child's developing cancer as a binomial spatially correlated random variable. We formulated the variogram for the risk in such a way that we could estimate it from the registry of cases held by the West Midlands Regional Health Authority, and we fitted a model to the estimates. Then, armed with the model and the data, we kriged the risk by binomial co-kriging. Somewhat to our surprise patches of greatest risk appeared in the countryside and suburbs rather than in the inner cities. You can find the relevant theory and results in Oliver *et al.* (1993) and Webster *et al.* (1994).

One feature of working with data of this kind is that there is a large error due to their binomial nature. In our study we had 840 electoral wards as the zones in which 605 cases of childhood cancer had been diagnosed in a period of 5 years. Even with this number of data the variogram was sensitive to small additions and removals of data, as when cases were reviewed and diagnoses changed by the oncologists. I am therefore concerned about the authors' analysis of the campylobacter data with only 234 cases among 248 postal zones. Can the authors tell us what the errors are and how these propagate into the estimates of the risk? I should like to see the variogram of the risk with error bounds on it.

I was puzzled by the variogram of the radioactivity. The authors present the experimental variogram. It is on a logarithmic scale, and so the fluctuation is enormous at lags beyond about 500 m. They then choose a Poisson model to describe the variation with a variogram that is several orders of magnitude larger over its whole length. The shape is different; it increases more markedly than the experimental variogram, and that is why the estimates vary more.

This brings me to the general matter of prediction errors. Geostatisticians have recognized for years that the kriging variances depend on the variogram. If the variogram is seriously in error, either because it is poorly estimated or because a poor model has been fitted to it, then the kriging variances will be poor estimates of the true prediction variances, i.e. mean-square differences between the true values and

the estimates, often denoted MSE for mean-square error. Further, for any one region practitioners usually use just one variogram, and so the kriging variances depend on the sampling configuration wherever it happens to be, and so they are rather general. In consequence, those trying to improve the technology now try to have independent sets of data for trial regions against which they can validate a new method. The variable of interest is kriged at the points in the independent set and the MSE is computed. This may be compared with the mean kriging variance. If the method is sound and the variogram has been satisfactorily estimated and modelled then the MSE and the mean kriging variance should be equal. More importantly, if the new method is to be an improvement then its MSE should be smaller than those of existing methods. The authors recognize this, but their standard kriging of radioactivity was done with the variogram of equation (23) and $\delta = 2$. Practitioners in the earth sciences know this as the Gaussian function. They also know to shun it because the kriging systems containing it are so unstable. This is most lucidly illustrated by Wackernagel (1995). The authors allow us only to compare maps of estimates visually and to accept the qualitative view of Dr Simon. So, can the authors tell us the errors in their two case-studies? What were the MSEs? Have they tried conventional kriging with a more stable well fitting variogram, and if so with what result? As a practitioner I should like to know that the MSEs using Markov chain Monte Carlo sampling are less than the best of current techniques if I am to add it to my repertoire.

I have provided a little extra background and raised enough doubts to debate for one evening. More positively, we in the earth sciences have a good grasp of linear geostatistics for straightforward estimation of ore grades, nutrient concentrations in soil and pollutants in the environment. When we want to use estimates for decision and combine them with other information we know that in many instances we must grapple with non-linearities for which our tools are inadequate or even lacking. So I am much encouraged to find professional statisticians such as Professor Diggle and his colleagues taking such a strong interest in these matters. They have shown us a promising new approach. I should like to think that they will join with earth scientists, get their hands dirty, both literally and figuratively, and solve a few more of our problems. We can provide them with plenty.

In conclusion, this has been an interesting and stimulating paper. On behalf of all present I thank you for it.

Andrew B. Lawson (*University of Abertay, Dundee*)

It is with great pleasure that I second the vote of thanks for this interesting contribution to the field of geostatistics. For a considerable time there has been little attempt to develop or generalize the methods that are implicit in the kriging literature, to situations where observational error is non-Gaussian. This is therefore a particularly welcome development.

As seconder of the vote of thanks, it is traditional to voice some (small) concerns about the work, and my comments are more of a comparative nature, emphasizing how the proposed methods could be matched by alternative (more Bayesian) approaches. The method proposed generalizes the Gaussian measurement error model in geostatistics, while keeping a spatial Gaussian process as the underlying model for the spatial correlation structure. The resulting *generalized linear* geostatistical model is sampled using Markov chain Monte Carlo methods. The Bayesian connections are particularly evident. Define the likelihood of data $\mathbf{y}$ given a random field $\mathbf{S}$ and parameters $\boldsymbol{\theta}$ as $L(\mathbf{y}|\mathbf{S}; \boldsymbol{\theta})$. Assume that both $\mathbf{S}$ and $\boldsymbol{\theta}$ have prior distributions $g_1(\mathbf{S}) \sim \mathrm{MVN}(\boldsymbol{\mu}, K)$ and $g_2(\boldsymbol{\theta})$, for the moment undefined. The full posterior distribution is proportional to $L(\mathbf{y}|\mathbf{S}; \boldsymbol{\theta})\, g_1(\mathbf{S})\, g_2(\boldsymbol{\theta})$. The parameter vector $\boldsymbol{\mu}$ can be specified to consist of spatial trend or other covariables, K is a positive definite covariance matrix which can be parameterized by using a specified covariance model. Warnes (1987) demonstrated that, if the likelihood was normal and $g_2(\boldsymbol{\theta})$ uniform, then undertaking maximum *a posteriori* (MAP) estimation for α in $\boldsymbol{\mu} = F\alpha$, where F is a design matrix, conditional on K, leads to universal kriging estimators. However, sampling of the full posterior distribution (including prior variation in the covariance parameters in K) will not lead to equivalent geostatistical estimators for the generalized linear models of the authors, unless MAP estimates are used as summaries. Posterior sampling of the above model has been examined by Lawson (1996, 1997), who also compared approximate MAP estimation with full posterior sampling summaries for a variety of likelihood models. The earliest use of approximate MAP estimates for a Poisson likelihood with specified covariance structure in a disease mapping application (putative health hazard) is found in Lawson (1994).

A second issue is the assessment of which type of spatial prior model should be used. Already a range of models has been proposed for this task from intrinsic autoregressions to types of conditional autoregressive priors (particularly, for example, in disease mapping; see for example Lawson and

Cressie (1998) and Lawson (1998). Can the authors specify *when* it would be appropriate to use parameterized covariance models of the geostatistical type as opposed to the other models, or indeed in the campylobacter example, as opposed to more parsimonious models for smoothing surfaces (e.g. kernel smoothing)?

The campylobacter example also raises some questions. The data are the addresses of cases of an infectious disease, which are likely to be best modelled as a space–time infection process and the use of a smoother for the 'declustered' spatial marginal of these data appears to be for convenience rather than a panacea: surely the original data should be described and not a 'declustering' of the data. Indeed what does a static image of an infectious process tell us about the progress of the infection process?

Finally I have some questions about the sampling methods used by the authors and sensitivity analysis.

(a) In the paper the surface $\mathbf{S}$ is sampled during iterations but is *estimated* in the final converged sample. Should this not have been averaged from posterior samples or conditionally sampled?
(b) The use of uniform priors for all parameters must have an effect on the sampled output. Do the authors have any comments about the sensitivity to changes of priors for parameters, and also do they have any idea what edge effects arise in these procedures and what, if any, effect do different priors have on these edge effects?

The vote of thanks was passed by acclamation.

Chris Glasbey, Graham Horgan and David Elston (*Biomathematics and Statistics Scotland, Edinburgh*)
The authors are to be congratulated on an elegant paper, which neatly combines kriging as weighted regression (Stein and Corsten, 1991) with recent approaches to generalized linear mixed models, again illustrates the power of Markov chain Monte Carlo methods, and is a further step towards a general spatiotemporal modelling framework. We see this paper as providing a very general methodology for modelling correlated data. However, we found the two examples somewhat disappointing: the absence of covariates makes the analyses seem simplistic, and the lack of display of the raw data hinders a critical assessment of the results.

Do the Rongelap data exhibit extra-Poisson variability? We ask, because the empirical variogram in Fig. 5 looks almost flat, which would imply that there is no spatial correlation in S and interpolation is unnecessary. Also in Fig. 5, we were surprised by the lack of agreement between the fitted and empirical variograms. It does not seem sufficient for the empirical variogram to be within the sampling envelope; should the fitted curve not be much closer when the same data have been used to estimate both variograms? The authors show the importance of including parameter uncertainty in inference about radiation maxima, but inference still rests on assumptions of stationarity and isotropy for S and a parametric model for ρ. Perhaps the paper could have gone further and included model uncertainty (Draper, 1995), though may be there is a limit to what can be done before credibility limits become arbitrarily large!

For the campylobacter data, it would have been interesting to see the results if temporal aspects had not been marginalized out of the analysis, and if rural postcodes had been modelled as area integrals rather than as point samples. We are pursuing both extensions in our work with digital images and spatiotemporal data. There are many opportunities for developing statistical methods for interpolating, smoothing and/or deconvolving images, by building on the ideas in this paper. Image data are usually area integrals and S often needs to be modelled by something akin to a piecewise constant, rather than a Gaussian, process. Also, we have used generalized additive models to fit spatiotemporal travelling waves to red grouse counts (Moss *et al.*, 1997). This methodology presents us with an opportunity to model the residual spatiotemporal correlation.

Gudmund Høst (*Norwegian Computing Center, Oslo*)
The authors are to be congratulated on a most stimulating paper on combining geostatistics with the flexibility of generalized linear models, as well as on the efficient implementation of their method through Bayesian computation.

A crucial question for the statistician when analysing spatial data is whether to model the full probability distribution or only to model statistical moments of low order. Kriging prediction is based on specified parametric forms for the mean and covariance (or variogram) functions only. In comparison, the authors go further and specify a wide class of parametric models for the underlying

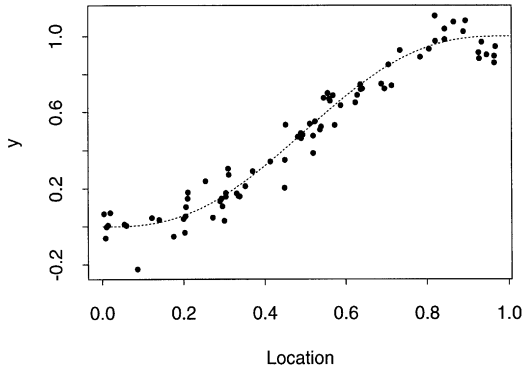


Fig. 13. Simulated data (•) and underlying trend function (·····) for a model with exponential covariance function with range 0.054, $\sigma = 0.1$ and $f(s) = 10s^3 - 15s^4 + 6s^5$

probability distribution of the data. This is necessary if the aim is to analyse extremes or excursion sets of the process.

Alternatively, the kriging approach of modelling the mean and covariance parametrically may be extended to a wide class by modelling these moments nonparametrically. This approach may be preferred in data-rich applications when the purpose is spatial prediction and assessment of prediction errors. In Høst (1996) a flexible framework for the prediction of a spatial process with unknown trend and correlated residuals is presented. For an example in one dimension, consider a continuous random process $y(s)$, where s is a location on the real line. Let $y(s)$ have the decomposition

$$y(s) = f(s) + v(s),$$

where $f(s)$ is a smooth trend function and $v(s)$ is a zero-mean, second-order stationary residual process.

In applications where the trend is unknown, it may be unrealistic to specify it parametrically. In particular, the kriging predictor will be biased under the model given above. Consequently, I suggest a local parametric approximation to the trend function f within a window of radius h and I derive an optimal predictor for this local model. The global properties of the predictor will be governed by h and a kernel function, and a framework is obtained in which both local polynomial regression estimation (Hastie and Loader, 1993; Fan and Gijbels, 1996) and kriging prediction can be described. In particular, I give an expression for the prediction error which also includes a bias term.

Fig. 13 shows simulated data from a process of the type described above, and Fig. 14 shows the estimated 95% prediction intervals for a local linear predictor and a linear trend kriging predictor. The bandwidth h in the predictor proposed was chosen by cross-validation. Fig. 14 indicates that kriging has

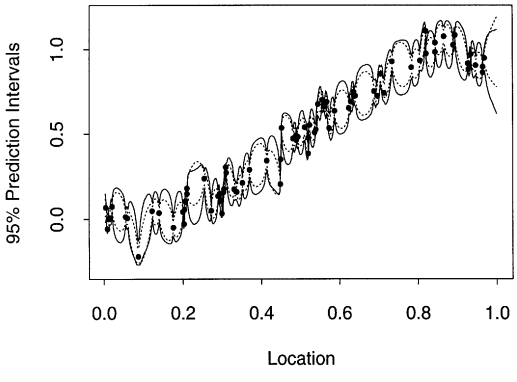


Fig. 14. Prediction intervals from local polynomial kriging (———) and universal kriging for the data of Fig. 13

smaller prediction errors than does the local polynomial predictor. However, cross-validation shows that the method proposed has both smaller prediction errors and more realistic estimates of these errors than does kriging. This is due to bias effects and to confounding of trend and residual structure in the kriging approach.

Future extensions of this approach may include nonparametric modelling of the spatial covariance function, possibly along the lines of Sampson and Guttorp (1992).

M. A. Mugglestone (*IACR–Rothamsted, Harpenden*) **and M. G. Kenward** (*University of Kent, Canterbury*)

We would like to pick up on the authors' observations that their regression parameters have conditional rather than marginal interpretations, and that the distinction between these two types of parameters is recognized for generalized linear modelling with longitudinal data. We are concerned with applications in which marginal interpretations of regression parameters are to be preferred, but where we do not wish to rule out conditional spatial prediction. The specific problem that we address is that of logistic modelling of spatially referenced, binary (presence–absence) data.

The regression problem may be tackled using generalized estimating equations (see, for example, Diggle *et al.* (1994)). In this framework, we can allow for spatial association through ratios of odds of presence (of a plant or animal species, or a disease) at different locations. There are several ways in which the behaviour of these odds ratios can be modelled as functions of spatial distance, by analogy with the variogram. This approach directly parallels the use of serial odds ratios in generalized estimating equations for longitudinal data (see, for example, Fitzmaurice and Lipsitz (1995)).

If, in addition to the estimated marginal regression model, conditional spatial predictions given neighbouring responses are required, then these can be calculated by combining marginal probabilities of presence from the logistic regression with a model connecting odds ratios of presence at different locations using iterative proportional fitting (Bishop *et al.*, 1975).

Clearly, our approach is not a general solution to the problem of marginal regression and spatial prediction with non-Gaussian data, and in this respect it is much more limited than the authors' conditional, Gaussian latent variable framework. However, we would argue that, for binary data, a measure of association based on odds ratios is more natural and easier to interpret than is the variogram.

Finally, we join the other discussants in congratulating the authors on a most interesting and stimulating paper.

John T. Kent (*University of Leeds*)

The basic idea of the paper, combining a generalized linear model for the observations with a spatial Gaussian process for the signal, is very elegant and powerful. My comments are directed at the simulated case-study, Section 6.1, which involves $n = 150$ equally spaced observations in $d = 1$ dimension with covariance function $\rho(u) = \exp(-|\alpha u|^{\delta})$, $0 < \delta \leqslant 2$.

Contrary to the claim in the paper, the covariance matrix Σ of the data is *not* singular for $\delta = 2$, merely very ill conditioned. To understand this result, it is easiest to switch to the spectral domain,

$$\rho(u) = \int \exp(\mathrm{i}u\omega) f(\omega)\, \mathrm{d}\omega,$$

with

$$\begin{aligned} f(\omega) &= C\alpha^{-1} \exp(-\omega^2/4\alpha^2) && (\delta = 2), \\ f(\omega) &\sim c_\delta \alpha^\delta |\omega|^{-1-\delta} && \text{as } |\omega| \to \infty \ (\delta < 2), \end{aligned}$$

exponential and power decay respectively, for large $|\omega|$ (e.g. Kent and Mardia (1994)).

The faster decay rate at $\delta = 2$ leads to a dramatic worsening of the condition number (the ratio of the largest to smallest eigenvalues of Σ) at $\delta = 2$ compared with δ near 2, e.g.

$$\begin{aligned} \delta = 1.95, &\quad \mathrm{CN} \approx 30\,000, \\ \delta = 2, &\quad \mathrm{CN} \approx 10^{60} \end{aligned}$$

(the latter figure is very approximate). The continuity properties of the process $X(t)$ also change abruptly with δ: continuous, not differentiable for $\delta < 2$; infinitely differentiable for $\delta = 2$. A broader range of continuity behaviour can be obtained with a class of covariance functions depending on Bessel functions, $\rho(u) = |u|^{\delta/2} K_{\delta/2}(|u|)$, $\delta > 0$ (e.g. Kent (1989)).

The parameter δ, $0 < \delta \leqslant 2$, determines the fractal dimension of the path of $X(t)$ by $D = 2 - \frac{1}{2}\delta$, through the behaviour of the covariance function near the origin, $\rho(u) = 1 - O(|u|^{\delta})$ as $u \to 0$. The estimation of δ requires an intricate statistical analysis, especially for δ near to 2 (Kent and Wood, 1997). The effect of δ is most noticeable at short lags of the process (equivalent to large frequency). Degrading the data by adding white noise to the process is equivalent to adding a constant term to the spectral density and can make the estimation of δ dramatically more difficult. Thus I would expect substantial problems in the estimation of δ and some confounding with the estimation of α. The examples of the paper seem to bear out these reservations.

A. Stein (*International Institute for Aerospace Survey and Earth Sciences, Enschede, and Wageningen Agricultural University*)

Let me first congratulate the authors on their very interesting paper. The link between spatial predictions and the generalized linear model, as well as use of Markov chain Monte Carlo methods, appears to be a great step forwards, despite some computational difficulties. In the two examples individual distributions of the data overcome the need to make an ergodicity assumption. There are, however, still some elements that intrigue me.

(a) First, the authors focus on spatial prediction, which, admittedly, is part of geostatistics. However, spatial sampling and the support size have not yet been addressed, and these appear to have a major implication on the grey tone maps (De Gruijter and ter Braak, 1990; Van Groenigen *et al.*, 1997). The most prominent boundary in Fig. 12, for example, i.e. the boundary in the north–south direction at an x-co-ordinate of 35300 with the log-odds ranging from -0.5 to 1.5 within a 4-km range, is seemingly caused by clustering of the data points and may be totally artificial. Also, the northern part of the region has not been sampled at all but still shows changes in grey tones. In contrast, the sparsely sampled part in the east shows no boundaries — but does seeing no boundaries mean that there are no boundaries?

(b) Model-based geostatistics appears to be extendable to multivariate predictions (Stein and Corsten, 1991). For p realizations of which the kth obeys the distribution $f^{(k)}(y^{(k)}|s)$ equation (12) changes to

$$f(y) = \int \prod_{k=1}^{p} \prod_{j=1}^{n_k} f_j^{(k)}(y_j^{(k)}|s_j)\, g_{n_1+\ldots+n_p}(s)\, \mathrm{d}s.$$

The problem then is to find a proper estimate for $g_{n_1+\ldots+n_p}(s)$, i.e. the multivariate probability density function for p properties. Moreover, stratification (or segmentation) of an area into homogeneous subareas makes the distributions for each individual location depend on the stratum where that observation occurs (e.g. Stein *et al.* (1991)). Do the authors agree with these points?

(c) Many environmental data, such as for heavy metals and in precision farming, show severe non-stationarity, like trends or hot spots. This can be treated with increments. Could the authors comment on how increments would fit into model-based geostatistics?

(d) Finally I wondered why the authors did not present the kriging variance for any of the examples.

Peter Clifford (*University of Oxford*)

I have a small concern about what is actually being measured when we look at the campylobacter data. The parameter p of the binomial distribution corresponds to the proportion of campylobacter out of three types. This proportion will be affected by all three components. Which effect are we observing?

Secondly, I would like to see some sort of cross-validation to assess the model fit. At the expense of computer time it is possible to carry out statistical analyses, missing one observation out at a time, and then to see what would be predicted for the missing value. Perhaps this is something that would be useful to do.

Finally, I am concerned about the way in which computer-intensive statistical analyses are presented. I would like to argue that

$$\text{MCMC} = \text{mediæval mathematics}$$

In the early days, when a mathematician discovered a new result it would be distributed by correspondence as a statement of the theorem without proof. The techniques of proof were highly guarded

secrets. It was the job of lesser mortals to work out why the result was true. Markov chain Monte Carlo (MCMC) results are a little like this. Very distinguished researchers arrange that a complicated computer program is written (probably funded by a 3-year research grant); they run the program, and a histogram is produced. Our job is to see whether we believe what they have done. This is not easy.

Before computer-intensive methods came along, statistical calculations essentially consisted of substituting in a formula, although there might be some dispute over which formula to use. Everybody could compute and recompute the formula to check the answer, and the formula could be recycled to look at other sets of data. Now, when the formula has been replaced by a computer program, we are stuck. Are we to believe that all the bugs have been removed from the computer program? Maybe there was a slip in typing in the data. Did last Thursday's computer run become mixed with Tuesday's? It is impossible to say. We must just believe that the distinguished authors have got it right, perhaps uneasily noting

> In the development of our algorithm, we often found apparent convergence of the routine to a false equilibrium. From our experience we would strongly advise the users of MCMC methods to calibrate and test routines carefully before applying them.

I agree. The reality is that different programs and different programmers can produce different answers. We must face up to the fact that a *proof* is needed. This means that people must be willing to put their programs on the Internet, so that others can see the secrets and try the methods for themselves on their own data sets.

Anthony W. Ledford and Pauk K. Marriott (*University of Surrey, Guildford*)

The paper's authors are to be congratulated for developing a widely applicable methodology that does not rely primarily on Gaussian assumptions. At a secondary level, however, Gaussianity of the process $S(\mathbf{x})$ is assumed. In practice, non-Gaussian spatial processes are often encountered, e.g. in turbulent fluid flow (She, 1991). Our findings suggest that assuming $S(\mathbf{x})$ to be a Gaussian process may lead to difficulties when extreme values issues are important. We illustrate this in the context of the Rongelap example.

Following Section 6.2, suppose that Y_i is Poisson distributed with mean $\lambda(\mathbf{x}_i) = \exp\{\beta + S(\mathbf{x}_i)\}$ where $S(\mathbf{x})$ is a heavy-tailed stationary process with symmetric marginal distributions. Biased values of $\lambda(\mathbf{x}_i)$ may be obtained if $S(\mathbf{x})$ is modelled as Gaussian, as overprediction and underprediction will occur at regions of low and high radioactivity respectively. Similarly, biases may result if $S(\mathbf{x})$ is modelled as Gaussian when in fact it is light tailed. If interest is in the typical levels of radioactivity, then these biases may have a small effect overall. However, as extreme levels are of interest here, the effect may be large because small relative biases in $\lambda(\mathbf{x}_i)$ can lead to large relative biases in the probabilities associated with extreme observations. We now examine this in greater detail.

Let $Y_{\lambda\epsilon}$ be a Poisson random variable with mean $\lambda(1+\epsilon)$, $\bar{F}_{\lambda\epsilon}(y) = \Pr(Y_{\lambda\epsilon} > y)$ and

$$U(y;\, \lambda,\, \epsilon) = 100\{\bar{F}_{\lambda\epsilon}(y) - \bar{F}_{\lambda 0}(y)\}/\bar{F}_{\lambda 0}(y).$$

Illustrative values of $U(y;\, \lambda,\, \epsilon)$ are shown in Table 1. These suggest that small relative errors in λ yield larger relative errors in the extreme probabilities, and that a given (fixed) relative error in λ has an increasing effect as more extreme levels are considered. Since extreme values are important here, the assumption that $S(\mathbf{x})$ is Gaussian needs careful verification. More generally, a misspecification of the link function or omission of covariates may lead to similar difficulties.

Table 1. Illustrative values of $U(y;\, \lambda,\, \epsilon)$

λ	100ϵ	*Results for the following values of y:*					
		10	*12*	*15*	*19*	*24*	*30*
5	1.0	6.8	8.8	12	16	22	29
	2.0	14	18	25	35	49	66
7	1.0	5.1	7.0	10	14	20	27
	2.0	10	14	21	30	44	62

Our questions are these: can the authors provide diagnostics for checking the various components of the model and specifically whether Gaussianity of $S(\mathbf{x})$ is appropriate, can they offer guidance on how to proceed in cases where Gaussianity is inappropriate and do they envisage any other difficulties when their methods are used to obtain inferences concerning extreme values?

Murray Aitkin (*University of Newcastle*)
I have one comment on this interesting paper. The need for covariance structure modelling should be demonstrated by an areal map of residuals from the model *without* a structured covariance, apart from the simple random effect term. Such a map might also establish the type of dependence to be modelled.

A. C. Atkinson (*London School of Economics and Political Science*)
I have one point specific to this paper and one general point. The specific point is that, like some other contributors to the discussion, I am concerned that what is essentially a smoothing method was used in the estimation of *maximum* intensity. What can the authors say about the resulting biases?

The general point concerns colour contour plots. The proposer of the vote of thanks used a standard set of colours from seismology. This was not the same as the authors' set. Nor is it the same as those that have appeared in the statistical literature.

Colour plots are still absent from the Society's journals. Some are, however, to be found in the *Journal of Computational and Graphical Statistics*. For example, the colours used by O'Connell and Wolfinger (1997) to some extent follow those of atlases and so relate to something familiar. Going up takes us from blue to green to yellow to orange. But going 'below sea-level' the blue changes to purple then changes to scarlet, disconcertingly close to orange. Venables and Ripley (1994), p. 111, plot from blue to green to cream, whereas Becker and Cleveland (1996) plot ozone concentration from cerise through azure to blue.

Any one of these, and several other, schemes is comprehensible on its own. The problem arises when plots in different schemes must be compared. May I, through the authors, ask that the Society pay some attention to this obstacle to the clear communication of an understanding of data and their analyses?

Michael Boskov (*City University, London*)
I found the way in which this paper brought together the ideas of kriging and Markov chain Monte Carlo (MCMC) methods very interesting. I have been using MCMC methods for spatial modelling for a long time, although all my work has been based on the approach of Besag *et al.* (1991), in which hazard rates for areas are modelled as a Markov random field.

I have been applying this type of spatial modelling to the problem of premium rating by geographical area as described in Boskov and Verrall (1994). In this application as in the application here to the Rongelap Island data, a potential problem for both kriging and the simple Markov random field methods is the tendency to flatten peaks and troughs in the data. In the case of the Markov random field approach it may be possible to reduce this effect by using higher order fields as shown, for different applications, in Besag and Tjelmeland (1996).

With kriging it seems more difficult to develop models that are analogous to higher order Markov random fields. However, although kriging is sometimes contrasted with smoothing spline models, it has been shown in Cressie (1991) that a simple thin plate spline model, based on penalizing the likelihood by the integral of the square of second-order differentials, is equivalent to kriging with a variogram from a specific family of functions.

This special case of kriging and smoothing splines can be generalized within the context of kriging by using different forms of variogram. It may be, however, that an alternative generalization within the framework of smoothing splines, using higher order derivatives, would reduce the tendency to flatten the observed values while still eliminating much of the random noise obscuring the true pattern of variation.

Julia Kelsall (*Lancaster University*) **and Jon Wakefield** (*Imperial College School of Medicine at St Mary's, London*)
We enjoyed reading the paper and would like to describe how a spatial Gaussian process may be used to model risk surfaces in disease mapping studies.

The analysis of disease mapping data is based on the following hierarchical model (see, for example, Mollié (1996)).

Stage 1
Suppose that a study region A is partitioned into areas A_i, $i = 1, \ldots, n$, and let Y_i and E_i denote the observed and expected disease counts in area i respectively. We then have

$$Y_i \sim \text{Poisson}(\lambda_i E_i)$$

where λ_i represents the relative risk of area A_i.

Stage 2
Here a distribution is specified for the collection of relative risks $(\lambda_1, \ldots, \lambda_n)$.

Various models have been proposed for this distribution. A simple model assumes that the λ_i are independently and identically distributed from, for example, a gamma or log-normal distribution. However, relative risks at locations close together tend to be more similar than those which are further apart, and in this case the independence assumption is not appropriate. The conventional approach to overcoming this difficulty is to assume a Markov random field model which specifies the joint distribution via the conditional distributions (e.g. Clayton and Kaldor (1987) and Besag *et al.* (1991)). For example, let $R_i = \log(\lambda_i)$ and consider the simple conditional autoregressive model $R_i | R_{-i} \sim N(\bar{R}_i, \sigma^2/m_i)$ where $R_{-i} = (R_1, \ldots, R_{i-1}, R_{i+1}, \ldots, R_n)$ and $\bar{R}_i$ is the mean of, and m_i the number of, *neighbours* of R_i. A disadvantage of this model is that it was originally developed for regular lattices and consequently takes no account of the differing shapes and sizes of the areas. This has been recognized and some alternatives (e.g. Cressie and Chan (1989)) have been proposed, though all are based on the idea of modelling the joint distribution through the conditional distributions via consideration of the neighbourhood structure.

We have developed an alternative approach (Kelsall and Wakefield, 1997), in which we model directly the underlying continuous risk surface and from this derive the distribution of the R_i. Let $R(x)$ denote the logarithm of the relative risk surface. We then make the assumption that $R(x)$ is a realization of a stationary Gaussian process. In this case we may evaluate

$$R_i = |A_i|^{-1} \int_{A_i} R(x)\, \mathrm{d}x,$$

and the joint distribution of $(R_1, \ldots, R_n)$ follows from standard properties of the multivariate normal distribution. Implementation of this approach through Markov chain Monte Carlo sampling is straightforward.

When the conventional conditional modelling approach is used the resultant risk estimates are constant within each area. A major advantage of our formulation is that a posterior distribution for the underlying risk surface over the whole region is also produced.

The following contributions were received in writing after the meeting.

Adrian Bowman (*University of Glasgow*)
I would like to congratulate the authors on providing a very useful framework within which spatial inference and prediction can be performed on a variety of data structures. As the authors point out, kriging can be viewed simply as a form of local smoothing. This raises the issue of how the authors' proposals might apply to more general smoothing problems. In particular, generalized additive models (Hastie and Tibshirani, 1990; Green and Silverman, 1994) are widely used to provide flexible regression models, in both spatial and non-spatial settings. Their ability to include both parametric and non-parametric terms is particularly attractive.

There are many broad points of connection with the present paper. For example, the parameters controlling the covariance structure of the surface $S(\mathbf{x})$ correspond to an assumed degree of smoothing in a nonparametric model. Since the authors' procedures have the attractive feature of incorporating uncertainty about these parameters into the final predictions, I would welcome any comments or guidance which they could give on the extent to which their methods might extend to more general nonparametric models.

Edward Casson (*Lancaster University*)
The authors comment that conventional geostatistical methodology is inappropriate for some applications because this technique assumes Gaussianity in the observations. Another example is

when the extremal behaviour is required for some process which has been recorded at several locations. To assume Gaussianity would bias estimates of extreme quantiles, so alternative distributional assumptions are regularly made. Problems are regularly encountered when investigating the spatial nature of the tail behaviour of some process:

(a) there is unlikely to be a set of measurements recorded at the exact location of interest;
(b) there may be few sufficiently extreme observations that satisfy asymptotic arguments required for unbiased, and informative, tail estimation. This may be due to the short period for which measurements are available from that particular measuring station or the problems associated with measuring such extreme data.

Suppose that observations have been recorded at several locations, and that spatial homogeneity of the process over that region may be assumed. As with any spatial model, a better estimate of the tail behaviour may be obtained by utilizing data at nearby locations. Investigations of the parameters which describe the tail of the process at each location by using only the data that are available at that site are likely to reveal smooth, but non-linear, variation in the parameters. By describing the spatial behaviour of the extremal process parameters using latent spatial processes of the form

$$h\{\cdot(\mathbf{x})\} = \mathbf{d}(\mathbf{x})^{\mathrm{T}}\beta + S(\mathbf{x})$$

following the authors' notation and employing similar Markov chain Monte Carlo techniques it becomes possible to model the parameters' spatial variation as something that it often appears to be—smooth but non-linear.

This approach enables a prediction of extreme quantiles of a process at locations where no data have been collected or for cases where a very high threshold is used. A simulation study in Casson and Coles (1997) revealed that, for sufficiently smooth processes, reasonably precise inferences on extreme quantiles can still be made at sites with missing data. This supports the use of spatial interpolation within this modelling framework.

A natural conclusion to make from this is that, by adapting techniques introduced by the authors, it is possible to understand the spatial characteristics of the tail behaviour of an environmental process with only a moderately dense spatial network of measuring stations.

Noel Cressie (*Iowa State University, Ames*)

In this paper, the authors have presented a methodology that is known elsewhere in the geostatistics literature as Bayesian kriging (Kitanidis (1986), Omré and Halvorsen (1989) and Cressie (1991), p. 171). They have noted the importance of measurement error in a proper formulation of kriging (optimal spatial prediction); earlier appearances can be found in Cressie (1988) and Cressie (1991), pages 127–130. The so-called nugget effect (discontinuity of the variogram at the origin) c_0 is actually made up of two components:

$$c_0 = c_{\mathrm{MS}} + c_{\mathrm{ME}},$$

the microscale variance c_{MS} and the measurement error variance c_{ME}. Classical geostatistics has implicitly assumed that $c_{\mathrm{ME}} = 0$, which is often not appropriate (Cressie, 1988). In the model-based geostatistics proposed here, the nugget effect decomposition is achieved through marginal and conditional distributions but the two notions of microscale variation and measurement error are still present in the problem. This paper assumes $c_{\mathrm{MS}} = 0$, at least in all the examples given; see the authors' equation (23). This deficiency needs to be addressed.

The problem with model-based geostatistics can be the models chosen. We need good diagnostics to reject inappropriate models in favour of more appropriate models. The diagnostic based on the variogram, as calculated in equation (19), is problematic. It is not purely a function of $\mathbf{u}$ since it depends on location $\mathbf{x}$ as well. Therefore, spatial averages are not equal to ensemble averages. When they are, it can be useful; in my view the theoretical model given in Fig. 5 provides a poor fit to the empirical variogram.

I do not want to leave the authors with a negative impression. I enjoyed reading about their use of Markov chain Monte Carlo methods to solve non-linear inference problems for hierarchical models. The next thing that one should do is to give up the Gaussian model for $S(\cdot)$, assumed at the second level of the hierarchy. Recently, Mark Kaiser and I have developed statistical methodology for spatial conjugate priors for exponential family data models, e.g. spatial beta priors for binomial data models. This research will appear elsewhere.

David G. T. Denison and Bani K. Mallick (*Imperial College of Science, Technology and Medicine, London*)

We focus on a problem for which conventional geostatistical methodology, extended in this paper, may be inadequate. Haslett and Raftery (1989) analysed wind speeds at 12 meteorological stations in Ireland with the aim of predicting the mean output from a wind turbine generator over a long period of time. Before starting to operate a wind turbine at a given location (known as the prediction location) wind speed data at the prediction location are recorded for a short period of time. Thus, to test proposed methodologies we take a short run of data at one of the stations and try to predict its mean speed given the data at the other 11 stations.

Unfortunately, the usual assumption that correlations between stations are related (non-linearly) to the distance between them (equation (23)) was found to be inadequate for this data set. In fact, in Haslett and Raftery's (1989) analysis, the data at one location are completely disregarded owing to their deviation from the restrictive parametric assumptions and there is some evidence that the stations on the coast have a different spatial covariance structure from those inland (Guttorp and Sampson, 1989).

We use the Bayesian multivariate adaptive regression spline (BMARS) algorithm (Denison *et al.*, 1998) and the short run of data to find a nonparametric model for the relationship between the wind speeds at the prediction location and at the others, i.e. we estimate nonparametrically the regression function f which is given by

$$\mathbf{X}_p = f(\mathbf{X}_{-p}) + \text{error}$$

where $\mathbf{X}_p$ are the data at the prediction location and $\mathbf{X}_{-p}$ are the data at the other stations. Note that we may also jointly model f by using other covariate information. This allows us to find models for the wind speed at each location, including 'outlying' ones, as no assumptions of stationarity are made. For wind data, factors which can affect the correlation structure are the prevailing wind direction and the local topography of the site. A reanalysis of the Haslett and Raftery data set using the BMARS algorithm, and a similar data set of wind speeds in Crete, is the subject of on-going research by Denison, Mallick and Dellaportas.

With this work in mind I wonder whether the authors believe that there are any serious problems with assuming an isotropic correlation function in the real data analyses and whether there is any way to use covariate information to alter the correlation structure. For example, in the campylobacter infections data set might the correlations between data depend on whether the locations were both in an urban or rural area?

Philip Dixon (*Savannah River Ecology Laboratory, Aiken*) **and Marian Scott** (*University of Glasgow*)

We congratulate the authors on a useful extension of spatial models to non-Gaussian data. The Bayesian methodology also provides a natural way to account for the uncertainty in the variogram parameters and to estimate non-linear functionals (e.g. exceedance probabilities or high quantiles), which are often more relevant than the mean. Our detailed comments concern the Rongelap example.

Three details of the Rongelap model might repay closer scrutiny. The data are collected by *in situ* γ-ray spectrometry, so the recorded count is a spatial integral over an area with an effective radius of the order of tens of metres. When sampling points are closer than the effective radius, nearby observations are more highly correlated than are caesium concentrations at those points. Hence, the short-range spatial correlation is inflated. The statistical models describe counts, but the questions concern soil caesium concentrations, which are proportional to the count rate (counts per unit time) minus the background contribution. The background adjustment introduces extra-Poisson variation, because background count rates are random and must be estimated. This is especially important for locations with low caesium concentrations. The nugget variance is explained as measurement error and Poisson sampling variation. The replicability of concentration estimates is quite high, especially if the count times are long. We suspect that the nugget includes a substantial component of small scale spatial variation. Although these non-linear spatial models may be sensitive to model assumptions, it is not clear whether these three details affect the conclusions.

One practical difference between these results and previous results from kriging log-transformed concentrations is the smoothness of the predicted surface. How much of that difference is due to the difference in method and how much might be due to using a different variogram (e.g. choosing the empirical or Poisson variogram in Fig. 5)? How sensitive are predictions to the choice of prior, especially for variogram parameters? If the likelihood dominates the prior, reasonable choices of prior have little effect. This may not happen here because the variogram is imprecise. Our one technical

concern is the specification of uniform priors in a non-linear model. The likelihood is invariant to the choice of parameterization, but a uniform prior is not because of the Jacobian associated with the transformation. Might Bates's technique of putting a prior on the expectation surface be useful here?

Timothy C. Haas (*University of Wisconsin, Milwaukee*)

The authors are to be congratulated for giving a Monte Carlo approach to the difficult problem of predicting a non-Gaussian spatial process. Perhaps Monte Carlo simulation is the future of statistical estimation and inference. I conjecture, however, that for some seemingly intractable multivariate distributions an Edgeworth-like expansion may be computationally feasible. As an illustration, I sketch a non-Monte-Carlo alternative to the authors' Markov chain Monte Carlo approach.

Let $h(\cdot, \boldsymbol{\beta})$ be a non-linear function with parameter vector $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^{\mathrm{T}} \boldsymbol{\beta}_2^{\mathrm{T}})^{\mathrm{T}}$. Let $Y(\mathbf{x})$ be a Gaussian spatial process with covariogram $C(||\mathbf{x}_1 - \mathbf{x}_2||, \gamma)$ wherein γ are the covariogram parameters. Let $Z(\mathbf{x}) = h\{\mathbf{d}(\mathbf{x})^{\mathrm{T}} \boldsymbol{\beta}_1 + \mathbf{Y}(\mathbf{x}), \boldsymbol{\beta}_2\}$.

Step 1: find the first six multivariate cumulants of the random vector $\mathbf{Z} \equiv (Z(\mathbf{x}_1), \ldots, Z(\mathbf{x}_n))^{\mathrm{T}}$ (Finney, 1963).

Step 2: use these multivariate cumulants to compute the Gram–Charlier expansion of the multivariate distribution of $\mathbf{Z}$ (see Finney (1963) and Tan (1980)) in terms of $\boldsymbol{\beta}$ and γ.

Step 3: use maximum likelihood (ML) to estimate $(\boldsymbol{\beta}, \gamma)$, and then write down the ML estimates of the joint density functions $f_Z(\mathbf{z})$ and $f_{Z(\mathbf{x}_0),\mathbf{Z}}\{z(\mathbf{x}_0), \mathbf{z}\}$. Parameter uncertainty can be assessed with the Hessian matrix evaluated at the solution point.

Step 4: using a quadrature-based numerical integration routine, compute an approximation to (say) $E[Z(\mathbf{x}_0)|\mathbf{Z}]$ by using the densities found in step 3. Note that quadrature-based numerical integration of a one-dimensional integral is fast, efficient and numerically well understood.

Other than those that can be expressed as transformations to the multivariate Gaussian characteristic generating function (Finney, 1962) for what transformations $h(\cdot)$ can the multivariate cumulants of $\mathbf{Z}$ be found?

The asymmetric *linex* loss function of Zellner (see Cressie (1993)) can be used instead of squared error loss and, for the two examples considered by the authors, may be more appropriate.

Mark S. Handcock (*Pennsylvania State University, University Park*)

The authors are to be congratulated on developing a natural extension of both generalized linear and spatial models. Their approach signals a major advance in the models that are available to practitioners. The paper is especially noteworthy as it displays a major contribution of Markov chain Monte Carlo methods: they allow the scientist to focus on developing more realistic models with less deference to inferential concerns. My first point concerns the choice of models for $S(\mathbf{x})$. Model (23) is quite restrictive as it automatically excludes realizations that are differentiable. This seems unnecessary as classes of models with the same number of parameters exist that do not have this restriction (Handcock and Stein, 1993). Consider the following example of spring temperatures for a region of the northern Great Plains considered in Handcock and Wallis (1994). Fig. 15(a) shows the posterior densities for the smoothness (θ_2) of $S(\mathbf{x})$ under both the Matérn class and model (23). The smoothness is the fractional number of derivatives — $\theta_2 = \delta/2$ in model (23) — and is graphed on the scale $\theta/(1 + \theta)$. The prior for the range is uniform and the smoothness is uniform on the graphical scale. This reflects the belief that very smooth $S(\mathbf{x})$ are *a priori* less likely. The two model classes coincide at a smoothness of $\frac{1}{2}$. Under the Matérn class, the probability that $S(\mathbf{x})$ is at least differentiable is about 44%. To compare the shape of the densities the posterior for model (23) has been scaled to have the same mass below 1 as the Matérn class. In this region of comparability the shapes are similar. The major difference is that class (23) places zero mass on differentiable realizations.

Fig. 15(b) shows the posterior prediction variances for a location in the centre of the region. As expected the prediction uncertainty decreases as the perceived smoothness of $S(\mathbf{x})$ increases and is smallest for values excluded by model (23). Thus unless differentiable modes of variation for $S(\mathbf{x})$ can be excluded *a priori* this model class will be inappropriate.

My second point is a call for the routine use of such graphical model diagnostics for these models. Likelihood and posterior surfaces such as those advocated in Handcock and Stein (1993) and Handcock *et al.* (1994) are essential tools for exploring model idiosyncrasies and finding model misspecifications. When the data only provide modest information about the parameters, the understanding of their joint dependences is essential.

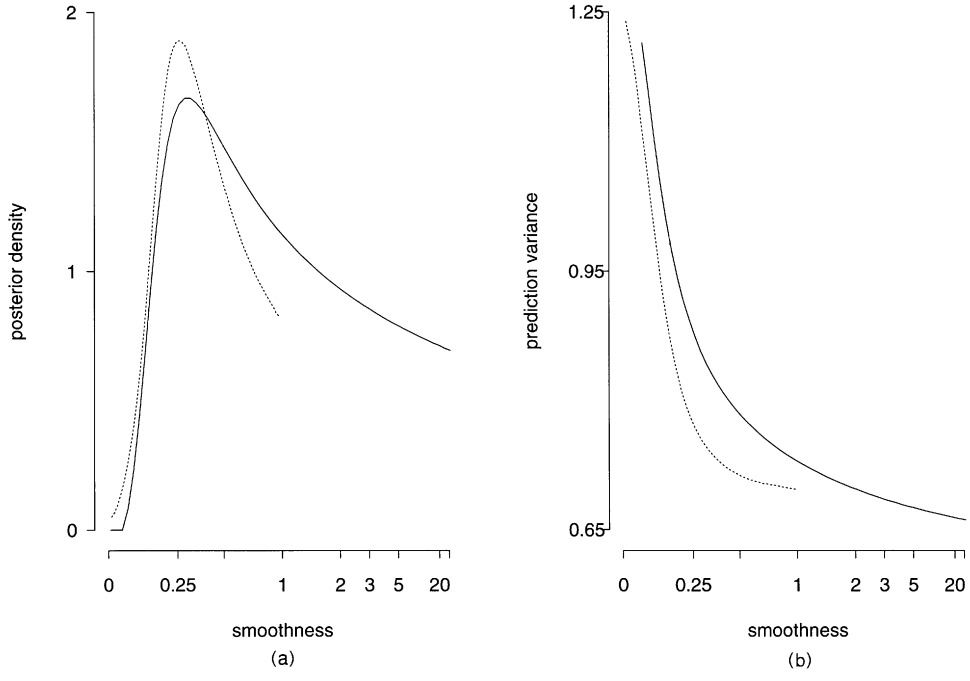


Fig. 15. (a) Posterior density and (b) prediction variance (·····, equation (23); ———, Matérn class)

C. C. Holmes and B. K. Mallick (*Imperial College of Science, Technology and Medicine, London*) Peter Diggle and his colleagues present a most interesting approach to geostatistical modelling. They adopt a Bayesian framework using kriging methods with a covariance function of the form $\rho(\mathbf{x} - \mathbf{x}') = \sigma^2 \exp\{-(\alpha u)^\delta\}$, where u is the distance between points. We would like to address some of the hidden assumptions in using a smoother of this form.

Firstly one should note that the covariance function is radially symmetric. One might say that it does not know its east and west from its north and south. This amounts to strong prior information. It is not entirely unfeasible in the Rongelap example that prevailing winds might cause stronger correlations among points lying in particular directions. It is straightforward to accommodate this possibility by taking α to be two dimensional and rewriting the correlation function between points $(\mathbf{x}, \mathbf{x}')$ as

$$\rho(\mathbf{x} - \mathbf{x}') = \exp\left\{ -\sum_{i=1}^{2} \alpha_i (|x_i - x_i'|)^\delta \right\}.$$

For fixed δ we can consider α_i^{-1} as a characteristic length scale for direction i over which the signal is expected to vary significantly. By comparing the joint posterior density of α_1 and α_2 we can infer whether the underlying process is indeed invariant to direction.

Secondly, the use of a kriging approach to surface fitting assumes that a constant level of smoothing is required over the design. This is implicit in the use of a stationary covariance matrix that is independent of location. For many problems this is not a realistic assumption. Greater confidence in the validity of this statement could be gained by partitioning the data into two or more spatially similar sets and performing inference on these sets independently. Having done so, posterior densities of the covariance parameters can be compared for universal convergence. If this analysis indicates that a spatially varying correlation is present then we might turn to other more adaptive methods that can readily accommodate this property. Recently Holmes and Mallick (1997) applied a Bayesian analysis to the use of radial basis functions. By assuming that the number and location of basis functions is unknown the model can perform local degrees of smoothing that is dictated by the data.

Finally we note what appears to be an omission in step 2(b) on p. 308. The proposal density

$p(S_i'|\mathbf{S}_{-i}, \boldsymbol{\theta})$ for this stage is not uniform or necessarily symmetric about the current state. Hence to maintain detailed balance a proposal ratio term, $p(S_i|\mathbf{S}_{-i}, \boldsymbol{\theta})/p(S_i'|\mathbf{S}_{-i}, \boldsymbol{\theta})$, is required in the acceptance probability $\Delta(S_i, S_i')$.

Geoff Laslett (*Commonwealth Scientific and Industrial Research Organisation, Clayton*)
The authors are to be congratulated on their innovative approach to non-Gaussian geostatistics. I suspect that most commentators will concentrate on matters of inference, so I shall discuss the data.

In my experience, applied geostatistics is a constant battle against data of dubious quality and representativeness. Here is a brief list of problems:

(a) typically 5% or 10% of mining samples are sent to a check laboratory, which systematically disagrees with the main laboratory;
(b) low and high response levels are measured by different procedures (e.g. copper assays);
(c) measurements are subjective (e.g. diamond valuations);
(d) instrumental drift can be severe;
(e) carry-over effects exist (e.g. soil acidity);
(f) inconsistent detection limits are applied (e.g. contaminated land studies);
(g) bulk surface samples are measured differently from borehole samples;
(h) measurements can depend on the time of day (e.g. ozone levels);
(i) refinements of the measurement procedure are confounded with spatial location.

Measurements of ^{137}Cs present their own problems. Caesium from nuclear fall-out is firmly adsorbed on the finer soil particles. The deposition of ^{137}Cs is fairly uniform, but with some spatial heterogeneity (Sutherland, 1994). In undisturbed soils, ^{137}Cs tends to be concentrated in the first few centimetres of the surface, with a roughly exponential decrease in concentration with increasing depth. Erosion by wind and water can severely affect the first few centimetres of soil, so that the caesium is redistributed along with the soil. Many studies (see Longmore *et al.* (1983) for an Australian example) have demonstrated that caesium depletion and build-up are linked to major areas of soil erosion and deposition. The ^{137}Cs concentrations can change rapidly over a few metres; slope gradient, drainage channels and man-made features such as fences can all be influential. A measurement on a steep slope may only weakly reflect the ^{137}Cs deposition at that location, let alone represent that of its neighbours 20 or 30 m away where the topography may be different.

Hence, it is not clear that the type of stochastic model considered in this paper is realistic for ^{137}Cs concentrations. Studies of ^{137}Cs should take into account topographic features and geomorphic processes (in particular, erosion); otherwise the results may be misleading. Have the authors done this for the Rongelap Island data?

Subhash Lele (*Johns Hopkins University, Baltimore*)
I would like to congratulate the authors for addressing the important problem of discrete data in space. These types of data occur commonly in ecological, environmental and public health studies. They also occur in medicine in the form of image data. This paper should help researchers in these fields substantially.

(a) I am usually faced with very large data sets where there are 4000 or more locations (Heagerty and Lele, 1998). We have developed an approach based on estimating functions to handle large data sets. Is the methodology in this paper computationally feasible in such situations?
(b) The convergence of Markov chain Monte Carlo methods is a generic problem. The authors use reparameterization and other tricks to overcome this problem. I wonder whether the inferences, particularly prediction intervals, are invariant to uniform priors on different parameterizations or not. If not, what is the authors' advice for users?
(c) How does the methodology in this paper compare with the EM, MCEM and MCNR algorithms (McCulloch, 1997)? These methods do not rely on the Bayesian approach and hence seem much more acceptable to scientists.

S. Nadarajah (*University of Plymouth*)
The authors are to be congratulated on a very interesting paper. Let me report on similar work that I am undertaking jointly with one of the authors of the paper.

There are many practical situations where interest is in the spatial variation of extreme values of a

variable. Examples include the dispersion of pollutants from a fixed source (such as a nuclear plant) and processes that lead to ignition. Consider the first example where we model the spatial variation by adapting the methodology of the paper as follows. Let $S(x_i)$ denote the signal at spatial location x_i and Y_i the excess of the pollutant's concentration over a high threshold at that location. We describe S by a Gaussian stochastic process with the correlation function given as in Section 6 of the paper and model $Y_i|S(x_i)$ by the reparameterized Pareto distribution:

$$f_i\{y|S(x_i)\} = \frac{1}{\sigma(1-\xi)}\left\{1 + \frac{\xi y}{\sigma(1-\xi)}\right\}^{-1/\xi - 1}$$

with the link function given by

$$M_i = E[Y_i|S(x_i)] = \sigma = \exp(s_i + \beta_0 + \beta_1 d_i),$$

where d_i is the vector distance from x_i to the source. The model (with five parameters in total) was fitted to simulated data using the Markov chain Monte Carlo methodology of the paper. Although we obtained reasonable estimates for ξ, β_0 and β_1, there was considerable difficulty in estimating the correlation function parameters α and δ. I believe that the authors encountered similar difficulties in their applications. Have they any solutions or suggestions for an alternative correlation function?

A. O'Hagan (*University of Nottingham*)

I congratulate the authors on their important and innovative paper. I have two related comments concerning their use of proper uniform priors and their Markov chain Monte Carlo algorithm.

When generating proposals for $\boldsymbol{\theta}$ and for $S(\mathbf{x}_i)$, they sample from the corresponding distributions. This will be inefficient if there is substantial information in the likelihood. In the case of $S(\mathbf{x}_i)$, is it possible to use conventional generalized linear model theory to obtain a useful normal approximation to the likelihood? If so, one could combine this with the prior to obtain a much better proposal distribution. Proposals would essentially then be derived from linear kriging.

Both $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$ are given proper uniform priors. In the examples, the authors do not specify the limits of these priors. Presumably they take account of the limits when sampling $\boldsymbol{\theta}$, but it is not clear that they do so when sampling $\boldsymbol{\beta}$-proposals, where a more sophisticated proposal distribution is described which should nevertheless be truncated.

In general, proper uniform priors may seem like a convenient way to avoid thinking about genuine prior knowledge, but they can lead to serious problems. In particular, when comparing models with different parameter sets, Bayes factors and the overall posterior distribution can be highly sensitive to the choice of limits for such priors. The authors do not explicitly consider model comparison or model averaging, but in their final paragraph there is a discussion of covariance models. If one wished to compare their covariance model with a model having a mixture form, expressing both short- and long-range dependence, then it would be necessary to think more carefully about prior distributions. One should either formulate proper prior beliefs or else apply a technique such as the fractional Bayes factor.

A. N. Pettitt and J. Hay (*Queensland University of Technology, Brisbane*)

We congratulate the authors on an excellent paper combining modern techniques and models to find answers for significant scientific problems.

We note some overlap with some current work. In Hay (1998), a generalized linear mixed model (GLMM) which involves a stationary autoregressive moving average model as the random effect is investigated using Markov chain Monte Carlo (MCMC) sampling and a Bayesian approach. In simulation studies, in style similar to those of the authors' Section 6.1, we found that the sweeping method (replacing β_0 by $\beta^* = \beta_0 + \bar{s}$) worked well for regression parameters, especially the intercept, but had little effect on parameters describing the random effects, which remained highly autocorrelated in MCMC runs. For our time series models when parameters approached boundaries which describe singularities, e.g. for the simple autoregressive model with the parameter going to 1 giving a random walk, chains were very slow to converge. In these cases posterior precision for the intercept in the linear predictor is small and a similar situation must occur for the spatial GLMM. Strategies such as fixing parameters as in Section 6.3 with $\delta = 1$ or choosing a different model need to be adopted. Sensitivity of inference to such choices needs to be investigated. As might be expected, models with more complex time series structure had better prediction and smoothing properties than models with more complex

fixed effects (more covariates), making the time series modelling worthwhile. We expect the same with spatial GLMMs.

In Weir and Pettitt (1997) we use a model that is similar to the authors' for binary data (absence or presence of animals in 10 km squares in Finland) but define the Gaussian spatial process on a lattice by its conditional autoregressive form involving nearest neighbours so that the full conditional distributions required for step 2 of the MCMC algorithm in Section 4 are straightforward and fast to compute.

Major computational effort must go into steps 2 and 4 in the MCMC algorithm because the inverse of the covariance matrix of S or (S^*, S) is not of a simple banded form and the full conditional distributions for S are not simple functions. For readers attempting similar analyses it would be useful for the authors to provide details of computing times.

Lastly, Pettitt and McBratney (1993) have discussed sampling designs for investigating variation at different scales of measurements. Such designs might be appropriate for investigations such as in the first example.

Sylvia Richardson (*Institut National de la Santé et de la Recherche Médicale, Villejuif*) **and Peter Green** (*University of Bristol*)

We congratulate the authors on a paper that provides further compelling evidence of the flexibility in both modelling and inference allowed by the Bayesian–Markov chain Monte Carlo partnership, this time in a class of spatial problems where the spatial scale is genuinely continuous.

However, we are puzzled by some aspects of the analysis in Section 6.3. The display of the prediction in Fig. 12 shows a very high degree of spatial smoothness. This is clearly a case where it would have been helpful to see some representation of the predictive variability, as was attempted in Section 6.2. The cases are distributed very irregularly, and we would have expected a realistic representation to have shown very large variability in that part of the region where the data are sparse. We note that this includes most of the region with high estimated log-odds.

Visually, Fig. 12 overstates the estimated spatial correlation (since we understand the units of measurement of length for which α is estimated as 6.5 to be approximately 60 km), but in turn the estimated correlations seem higher than are scientifically probable. Our partial understanding of the epidemiology of campylobacter infection points to localized contamination from animal or water sources, implying rather short-range spatial correlations. In thinking about the explanation for the inferences of high correlation presented in the paper, we surmise that there must be heterogeneity in spatial correlation between urban and rural areas, and that inference about α is being dominated by information from the urban cases. It would be interesting to extend the model to allow such heterogeneity, and to test our hypothesis.

Finally, we note that the approximate method of declustering that is used, which aims at focusing the analysis on outbreaks rather than cases, does not address the possibility of secondary infection, which may exhibit a different time pattern from multiple cases from the same external source; further, because of the differential sizes of unit postcode locations, we would expect some bias in the analysis caused by the different effect of the approximation in urban and rural areas, since non-related outbreaks at the same location will be more probable in urban than rural areas, other things being equal.

Michael Stein (*University of Chicago*)

Geostatistical practice would greatly advance by the widespread adoption of the methods described in this work. However, I have concerns about some of the details of this work. In particular, for the model (23) of the correlation function, when $\delta < 2$ the resulting process is not mean square differentiable and when $\delta = 2$ the process is analytic. Thus, this class of models does not include, for example, any models for processes that are once but not twice mean square differentiable. Handcock and Stein (1993) advocated the use of the Matérn model, which includes a parameter that allows for any degree of mean-square differentiability for the process. Its use requires the evaluation of modified Bessel functions, but this is not a serious obstacle to its adoption. As the paper notes, using $\delta = 2$ yields very nearly (but not exactly as the authors imply) singular correlation matrices, which is just as well, since a model that says that some process in space is analytic should never be used anyway.

Interpretations of posteriors may depend in important ways on the choice of parameterizations. For example, the authors note that the posterior means for α are much larger than the posterior modes in both examples. To see why the posterior for α can have such a fat upper tail, note that once α exceeds a certain value the observations are essentially uncorrelated so that further increases in α have almost no

effect on the correlation matrix of the observations. I suspect that the posterior means and modes of $\alpha' = 1/\alpha$ are in much better agreement than those of α in the examples. Incidentally, the fact that α' provides at least as natural a parameterization for the range of the correlation function as does α suggests that we might want to pay more attention to the priors than just choosing them to be uniform over a somewhat arbitrarily chosen region.

Finally, I must object to the claim in Section 6.2 that the empirical semivariogram is a consistent estimator of the true semivariogram. Under the natural asymptotic regime of taking increasingly more observations on Rongelap Island, the empirical semivariogram will at best be a consistent estimator of the behaviour of the semivariogram at the origin.

Dietrich Stoyan (*Freiberg University of Mining and Technology*)

In some geostatistical analyses we are close to point process statistics. This is the case when the locations $\mathbf{x}_i$ are irregularly or even randomly placed, as in example 2. Then it may be useful to apply ideas of point process statistics, and the geostatistical data can be interpreted as marked point process data, where the $\mathbf{x}_i$ are 'points' and the Y_i 'marks'. Though we have the same data structure as in geostatistics, it may be quite a different situation: it is not a random field that is observed (which has a value in each point $\mathbf{x}$ of the space) but only isolated points in which Y-values exist. A counterpart to the variogram of geostatistics is the mark variogram introduced in Gavrikov and Stoyan (1995). For the estimation of variograms it seems to be natural to use the estimator suggested in that paper or kernel estimators, instead of grouping the spatial separations.

A simple model of a marked point process is constructed by means of a point process N and an independent random field $\{Y(\mathbf{x}): \mathbf{x} \in \mathbb{R}^d\}$. The mark of point $\mathbf{x}_i$ of N is then simply $Y(\mathbf{x}_i)$. In this case the mark variogram coincides with the variogram of the random field, and still the points have the character of observation points as they clearly have in example 1.

However, more complicated models are possible and sometimes necessary, when the mutual positions of points have influence on the marks and the points are not neutral observation points. The local interaction of the points may counteract the random field correlation and thus a mark variogram is not necessarily negative definite, as shown in Wälder and Stoyan (1996), where points close together produce rather different marks.

It seems to me that example 2 is a point-process-related situation. The 248 locations are not independent observation points but locations, the relative positions of which may have influence on the observation results.

It would be interesting to see the empirical variogram (which I would interpret as a mark variogram) for example 2. Because of the particular form of local interaction of the points, I except here a negative definite variogram, which justifies the application of geostatistical methods.

C. K. I. Williams (*Aston University, Birmingham*)

The authors show how Gaussian random fields (GRFs) may be transformed through the link function of generalized linear models to provide useful geostatistical models. My comments pertain to the use of these kinds of model in the general prediction case, where $\mathbf{x}$ is a vector of covariates, not necessarily spatial in nature. Although GRF methods are used in some statistical fields, e.g. in computer experiments (see, for example, Sacks *et al.* (1989)), it is curious to me why they are not more widely employed. One early advocate of GRFs for general regression problems is O'Hagan (1978).

Several researchers including Wahba (1990) and Green and Silverman (1994) have shown how predictions may be made using a roughness penalty on what the authors call $S(\mathbf{x})$. However, these treatments typically use spline-type generalized covariances and do not deal fully with uncertainty in the parameters therein. Some recent work addressing these issues can be found in Barber and Williams (1997) and Neal (1997). Neal's treatment is similar to that of the authors, although he uses rather more sophisticated Markov chain Monte Carlo methods, including a hybrid Monte Carlo method for θ and Gilks and Wild's (1992) adaptive rejection sampling for the S_i. He applies the method to classification problems by using the logistic function (and its multiclass analogue) and to regression problems under the assumption of t-distributed noise. His code is available from

`http://www.cs.utoronto.ca/~radford`.

Barber and Williams (1997) provide a similar treatment for classification problems but use Laplace's method to approximate the integral in equation (10). Preliminary results indicate that the performance of these methods is competitive with other good classification algorithms.

In the general prediction problem, the use of an *isotropic* covariance function as is frequently used in

geostatistics may be unsuitable. One simple modification is to introduce parameters that allow separate scaling factors w_k on each input dimension, so that the distance between $\mathbf{x}$ and $\mathbf{x}'$ is defined by

$$d^2(\mathbf{x}, \mathbf{x}') = (\mathbf{x} - \mathbf{x}')^{\mathrm{T}} W(\mathbf{x} - \mathbf{x}'),$$

where W is a diagonal matrix with non-negative entries. This affords an implementation of the 'automatic relevance determination' idea of MacKay (1993) and Neal (1996). Assuming that the covariates are similarly scaled, a small value of the w_k-entry indicates that input k only gives rise to slow variation in the $S(\mathbf{x})$ surface. Results reported in Williams and Barber (1997) show that the posterior means of the ws can differ by several orders of magnitude in practice.

The **authors** replied later, in writing, as follows.

We thank all the discussants for the interest that they have taken in our work, and we are especially indebted to those discussants whose theoretical insights have clarified some problematic aspects of our modelling and inference. We have structured our reply around what seem to us to be the main themes which emerged during the discussion, and we apologize for any resulting omissions.

The model for the stationary Gaussian process

Our choice of the powered exponential family of covariance functions was, with hindsight, a bad one. Webster reminds us that practical geostatisticians know to avoid the correlation function $\rho(u) = \exp\{-(\alpha u)^2\}$, while Handcock, Kent and M. Stein give good theoretical reasons why this is so. We intend to rework our analyses using the Matérn class in place of the powered exponential. However, we feel that for data sets of the size, and sampling locations, of our two examples, it will be difficult to estimate more than one parameter in the correlation function. Hence, the choice of, for example, the shape parameter in the Matérn class is likely to be heavily influenced by the analyst's preconceptions (perhaps formalized through a prior distribution) of how smooth analytically the underlying surface is thought to be. We overcome this problem in the campylobacter example by fixing $\delta = 1$; a similar solution may help to resolve Nadarajah's problems as it does those of Pettitt and Hay.

Webster is concerned that by using the 'wrong' covariance function we may also produce the wrong prediction intervals. It is, of course, true that prediction using the wrong model may underestimate or overestimate prediction errors. However, treating an estimated covariance function as if it were known certainly underestimates prediction errors. We want our prediction errors to be small, but not spuriously small!

Bayesian inference

Cressie claims that our approach falls within the methodology known in the geostatistics literature as Bayesian kriging. We disagree, as that literature applies Bayesian methods only to the linear kriging class of models of Section 2.1 of our paper.

Several discussants (Lawson, Dixon and Scott, O'Hagan and M. Stein) commented that our use of proper uniform priors is not very natural, and at the meeting we were described as 'reluctant Bayesians'. This is fair comment. We were attracted to the Bayesian approach because it provides an elegant way of allowing for parameter uncertainty in setting prediction intervals, and to the Markov chain Monte Carlo (MCMC) machinery because it solves problems that are concerned with the prediction of non-linear functionals. We agree that vague prior knowledge need not, and probably should not, be expressed through uniform priors. We have some sympathy with Clifford's concerns about the proliferation of MCMC analyses whose results cannot easily be checked. Our view is that MCMC sampling is the technology of last resort, but it is nevertheless invaluable for the solution of otherwise intractable real problems. Unlike more standard solutions, MCMC procedures need extensive simulation testing for each different model fit. This detail is difficult to report concisely in print; but we aimed at least to report honestly how taxing this calibration problem is for our particular models and data sets.

The Markov chain Monte Carlo algorithm

Answering questions from Lawson and O'Hagan first, the estimated S is obtained by averaging samples once the chain is deemed to have converged, and sampling proposals for β were truncated in the algorithm. Holmes and Mallick ask about a possible omission in step 2(b) of the algorithm. The proposal density as specified in step 2 is univariate normal. Using this kernel, the acceptance probability is

$$\Delta(S_i, S_i') = \min\left\{\frac{p(\mathbf{Y}|\mathbf{S}, \boldsymbol{\beta})\, p(S_i|\mathbf{S}_{-i}, \boldsymbol{\theta})\, q(s_i, s_i')}{p(\mathbf{Y}|\mathbf{S}', \boldsymbol{\beta})\, p(S_i'|\mathbf{S}_{-i}, \boldsymbol{\theta})\, q(s_i', s_i)}, 1\right\}.$$

If $q(s_i, s_i') = p(S_i'|\mathbf{S}_{-i})$, then $\Delta(S_i, S_i')$ reduces to the expression given in the paper.

O'Hagan suggests that $\pi(S_i'|\mathbf{S}_{-i}, \mathbf{Y}, \boldsymbol{\theta}, \boldsymbol{\beta})$, in equation (16) of the paper, has an approximate normal distribution with mean and variance given by equations (3) and (4) of the paper. This approximate distribution can be used as the proposal distribution q. We believe that this could make our algorithm a little more efficient, but that it would not make a substantial difference to the results.

We are interested by the suggestions of Haas and Lele for developing less computer-intensive methods which avoid the need for the MCMC approach, and in particular we would like to explore the possibility of adapting some of the ideas in McCulloch (1997), although we are unclear how these methods will deal with parameter uncertainty. Maybe the ideal solution is to incorporate such techniques into an MCMC algorithm to provide better proposal distributions, similar to O'Hagan's suggestion, thus reducing the computational demands while retaining the inferential flexibility of the MCMC approach. However, this may still not entirely overcome the sample size restrictions of the current form of the algorithm, which can only comfortably handle data sets of size up to $n = 300$. To handle very much larger data sets, it may be necessary to implement either a localized version of the algorithm, in which updating considers only information from spatially close sites, or to use a Markov random field model for S. One example of this would be a 'conditional autoregressive prior' as mentioned by Lawson, although a difficulty with such models is that the interpretation of their parameters is specific to the configuration of sites on which they are defined.

The Rongelap data

Glasbey, Horgan and Elston were disappointed by the absence of covariates in the Rongelap Island data set, and Laslett gives us examples of potentially relevant covariates which might have been collected. But they were not. We strongly agree that the inclusion of relevant covariate information ('kriging with external drift', in conventional geostatistical language) can be of great value, both in reducing prediction errors and in scientific interpretation. Indeed, we regard the stationary Gaussian process S as being at least in part a surrogate for unidentified, spatially referenced covariates. For the Rongelap data, the original survey consisted only of the measurements at the 200 m grid spacing. An uninhabited, contaminated coral atoll is a hostile environment for fieldwork. We persuaded the investigators to return to the island to collect measurements from the four 5×5 grids at 40 m spacing, but this was the limit of what was feasible. We deliberately sited two of these 5×5 grids in the area of the island that was previously occupied by the islanders' dwellings and two in a previously unoccupied area, and we looked for differences between the responses in these two zones, but we found none and therefore treated the data as a single population for the subsequent analysis.

Glasbey, Horgan and Elston, and Aitkin ask whether it is even necessary to model spatial variation in these data above the Poisson variation of the counts. We did establish to our satisfaction that the data show very substantial extra-Poisson variation, and that this seems to be spatially structured. Fig. 16 shows a map of the signs of the residuals from a Poisson fit with constant intensity, i.e. a model which assumes that the counts $Y(\mathbf{x}_i)$ are mutually independent with $Y(\mathbf{x}_i) \sim \text{Poisson}(\lambda t_i)$. Similarly, Fig. 17 shows the map obtained with λ replaced by $\hat{\lambda}(\mathbf{x})$, given by Fig. 7. The majority of the spatial structure in Fig. 16 appears to have been removed in Fig. 17.

Dixon and Scott are right to point out that the subtraction of the estimated background radiation level from the raw count makes the Poisson assumption strictly an approximation, but we do not feel that this effect is sufficiently large to change our conclusions substantially. They also point out that the technology used to collect the caesium measurements itself induces short-range spatial correlation because the γ-ray camera integrates information from a circular zone centred on the measurement site. However, with a minimum distance of 40 m between measurement sites, this short-range effect is not sufficient to explain the observed spatial correlation structure of the data.

Webster and A. Stein ask about the prediction variance for $\lambda(\mathbf{x})$. Fig. 18 shows the mapped posterior standard deviation for $\lambda(\mathbf{x})$. This has quite a large amount of structure to it: spots of very small variation where the prediction locations overlap with data locations; largest variation on the edge of the island where the information from the data sites is least; lower than average variation over the regions covered by the four densely sampled grids. Clearly the sampling design has a substantial effect on this map, reflecting comments made by A. Stein and by Pettitt and Hay. Note also that the posterior

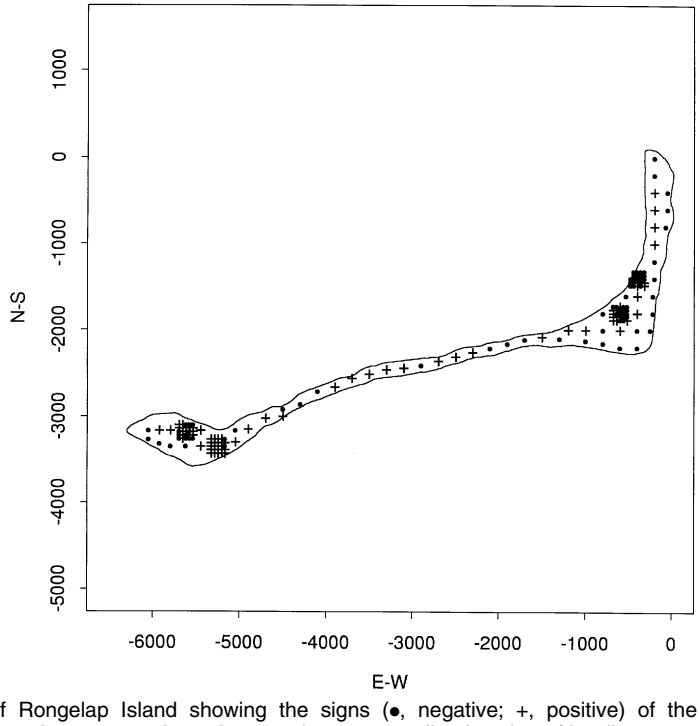


Fig. 16. Map of Rongelap Island showing the signs (•, negative; +, positive) of the residuals from an independent fit assuming constant intensity over the 157 sampling locations (the distance scale is as in Fig. 4)

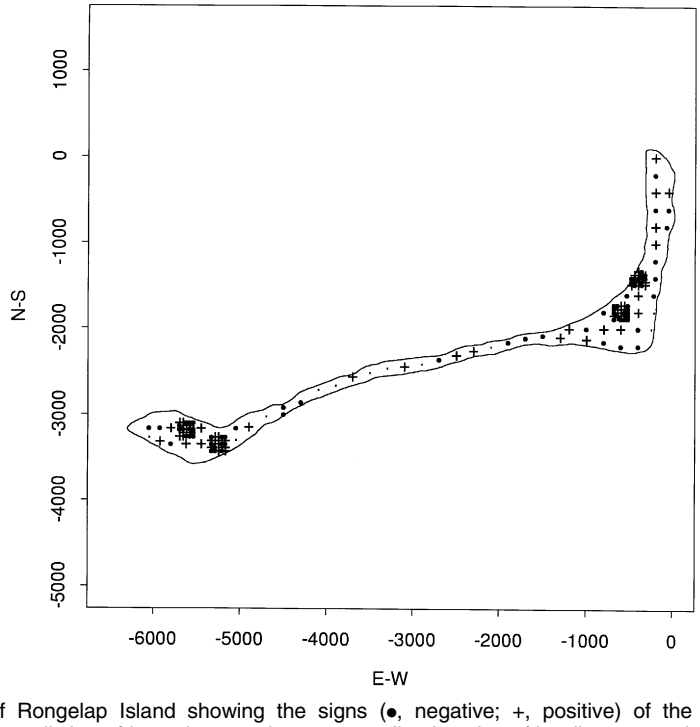


Fig. 17. Map of Rongelap Island showing the signs (•, negative; +, positive) of the residuals from the generalized linear prediction of intensity over the 157 sampling locations (the distance scale is as in Fig. 4)

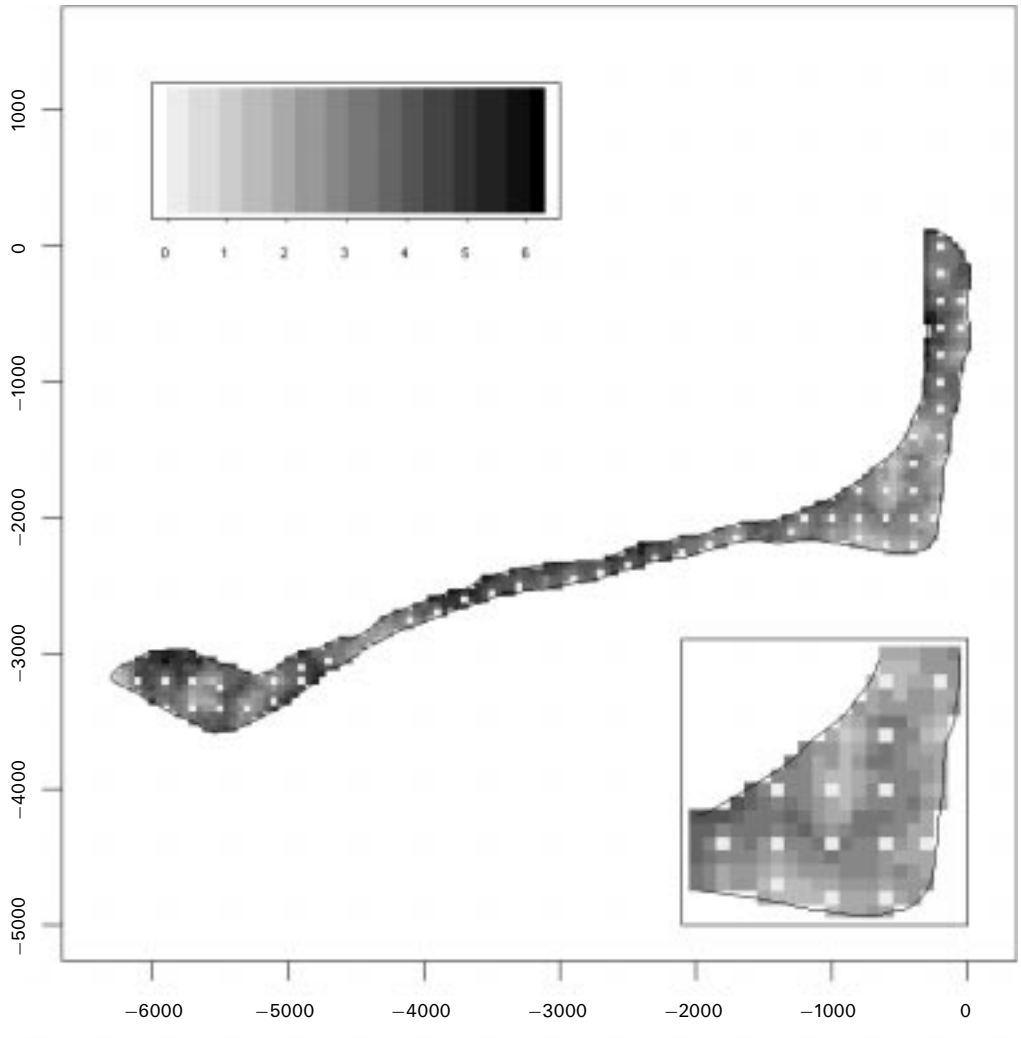


Fig. 18. Standard deviations of intensity of ^{137}Cs over Rongelap Island obtained by using generalized linear prediction (the distance scale is as in Fig. 4)

distribution of $\lambda(\mathbf{x})$ is highly skewed when $\lambda(\mathbf{x})$ is small, so the standard deviation is a very incomplete summary of the posterior variation.

The campylobacter data

We wanted to include the campylobacter example as an illustration of the scope for applying our methodology in an epidemiological setting, but we acknowledge that from a scientific point of view this particular application is somewhat unsatisfactory: the declustering of the data before modelling is somewhat *ad hoc*; the unit postcodes correspond to a very uneven spatial distribution over the study region, which includes both urban and extremely rural areas; and, as Richardson and Green point out, the disease is known to have a strong rural–urban risk gradient.

Despite our reservations about the analysis we do not share Clifford's worry about focusing on the conditional probability surface given the number of common enteric infections. This probability is directly related to the relative risk, which is of epidemiological interest in its own right. If the absolute

risk is also of interest, a more natural starting point might be to model each disease separately, e.g. as a pair of Poisson processes with respective intensities

$$\lambda_i(\mathbf{x}) = \lambda_0(\mathbf{x}) \exp\{S_i(\mathbf{x})\}, \qquad i = 1, 2,$$

in which $\{S_1(\mathbf{x}), S_2(\mathbf{x})\}$ is a bivariate stationary Gaussian process, $\lambda_0(\mathbf{x})$ is the spatial intensity of the population at risk and the Poisson processes are conditionally independent given $\{S_1(\mathbf{x}), S_2(\mathbf{x})\}$. However, the corresponding conditional model for the proportion of cases of type 1 would then be

$$\text{logit}\{P(x)\} = S_1(\mathbf{x}) - S_2(\mathbf{x}) = S(\mathbf{x}),$$

say. In this formulation, $S(\mathbf{x})$ is itself a stationary Gaussian process, so the two approaches are entirely consistent with each other.

As a general point we feel that it is reasonable to model risk as a continuously varying spatial process, and we disagree with Webster's comment that this is a 'non-problem'. However, it is certainly the case that, in applications of this kind, data on relevant risk factors are often only available in spatially aggregated form, as averages over discrete spatial regions. We are very much attracted by Kelsall and Wakefield's approach to problems of this kind. Their methodology may also be appropriate for the type of applications which Glasbey, Horgan and Elston discuss.

Diagnostic tools

The paper was rather thin on diagnostics, partly for brevity and partly as the data contain limited information which is difficult to assess owing to the spatial dependence. We have covered some additional aspects in the response above. We agree with Webster, Clifford, and Holmes and Mallick that cross-validation methods provide potentially powerful tools for identifying lack of fit with respect to a range of model features but, primarily for computational reasons, we did not undertake these. Such approaches may have clarified the effect of urban–rural heterogeneity in the campylobacter example. Model choice aspects, raised by Lawson and by Glasbey, Horgan and Elston, are important generally for spatial data but scope for their effective use in particular applications is limited by the amount of data that are available. We agree with Mugglestone and Kenward that special diagnostics are required for spatial non-Gaussian data, particularly when the observed data are highly discretized.

Extensions to our model

Cressie comments on potential extensions of our approach to allow for a non-Gaussian process S. We agree that this is an open and important problem. However, there is also scope for incorporating additional flexibility into the present model while retaining a spatial Gaussian process (univariate or multivariate) to induce spatial dependence. Several suggestions were raised in the discussion which can be handled this way. For example, Cressie points out that the nugget effect is made up of two components: a microscale variance and a measurement error variance. This ties in with Dixon and Scott's observation that the technology used to collect the caesium measurements itself induces short-range spatial correlation. In principle, our methodology could incorporate this type of contextual information into a more sophisticated two-scale model for the process S as mentioned in Section 7.

Holmes and Mallick, and Williams commented on our using only isotropic correlation structures. We do not believe that there is sufficient information in either of our specific applications to identify anisotropy, although we accept that in general it is useful to have the option of modelling directional effects. We also note that although we 'cannot tell our north–south from our east–west' these discussants all suggest correlation functions which depend on the potentially arbitrary labelling of north. Williams's distance measure but with W a general symmetric matrix seems to be more natural. Denison and Mallick discuss extensions to non-stationary S. Within our framework we can achieve similar flexibility, for example letting σ vary spatially by introducing an additional layer into the hierarchical structure, and making $\rho(\cdot)$ dependent on location and direction through a more general parametric model. For example, in related work concerned with modelling the space–time variation in rainfall intensity, we are thinking about ways of modelling non-stationary and non-isotropic spatial correlation with the local 'principal direction' of correlation determined by covariate information on wind speed and direction.

Extensions beyond our generalized linear set-up are outlined by Casson and by Nadarajah for estimating spatial variation in marginal parameters of extreme value analysis. This seems a valuable application which could potentially be widened to modelling spatial dependence of extreme value data by allowing S to vary over realizations in time.

Stoyan raises a different kind of extension, to marked point processes. Their essential feature is that the sampling locations (points of the process) cannot be assumed to be chosen independently of the real-valued process S (the mark process). We suspect that stochastic dependence between sampling locations and associated measurements is often present, but ignored, in standard geostatistical applications. For example, in sequential exploration of a potential oil-field, it would presumably be sensible to site new test drillings in areas which appear, on the basis of results from existing sites, likely to yield a positive outcome, but this would invalidate standard geostatistical inference. Stoyan's references point out some of the dangers of ignoring this kind of stochastic dependence.

While agreeing that all these possible extensions of our basic model structure are worth pursuing, we feel that the generalized linear setting with a Gaussian S is sufficiently general to be of interest in its own right.

Spatial smoothing

Atkinson and Boskov commented on spatial smoothing being used to estimate the extremes of the surface, postulating that the likely outcome is that the extremes are oversmoothed. Our approach avoids this problem and provides extrapolation beyond the observed range of the data, but at the cost of explicit modelling assumptions. For example, ordinary linear kriging with a particular assumed form of covariance function can be formally interpreted as a spline smoother (but note the discussion of this connection in Laslett (1994)). In our case the largest observed intensity was 15, which is oversmoothed in Fig. 6, but not in Fig. 7. This is further emphasized by Fig. 9(a) which shows that the maximum observed intensity falls in the lower tail of the distribution of the predicted spatial maximum. However, this extrapolation of the surface may be highly sensitive to distributional assumptions, as illustrated by Ledford and Marriott. This is particularly relevant given Dixon and Scott's reminder about the Poisson distribution being an approximation in the Rongelap application.

Other spatial smoothers were proposed by the discussants. Generalized additive models embody the extension of linear smoothers to the generalized linear setting. Høst, Bowman and Williams develop these connections. The benefits of the flexibility of Bayesian nonparametric smoothing methods are discussed by Denison and Mallick and by Holmes and Mallick.

These comparisons raise the question of whether it is better on balance to have the flexibility provided by nonparametric approaches or the parsimony and interpretability provided by parametric modelling. Operationally, both approaches produce estimated surfaces $\hat{S}(\mathbf{x})$ which are nonparametric in the sense that $\hat{S}(\mathbf{x})$ (as opposed to the model which leads to it) are not contained within a finite parameter class of functions. We prefer the parametric modelling approach presented in the paper because the degree of smoothing is then determined by the model fitted to the data, which in turn is determined according to a well-established (albeit highly non-automatic) general inferential paradigm.

References in the discussion

Barber, D. and Williams, C. K. I. (1997) Gaussian processes for Bayesian classification via hybrid Monte Carlo. In *Advances in Neural Information Processing Systems* (eds M. C. Mozer, M. I. Jordan and T. Petsche), vol. 9. Cambridge: MIT Press.

Becker, R. A. and Cleveland, W. S. (1996) *S-PLUS Trellis Graphics User's Manual*. Seattle: Mathsoft.

Besag, J., York, J. and Mollié, A. (1991) Bayesian image restoration with two applications in spatial statistics. *Ann. Inst. Statist. Math.*, **43**, 1–59.

Bishop, Y. M. M., Fienberg, S. E. and Holland, P. W. (1993) *Discrete Multivariate Analysis: Theory and Practice*. Cambridge: MIT Press.

Boskov, M. and Verrall, R. J. (1994) Premium rating by geographic area using spatial models. *Astin Bull.*, **24**, no. 1.

Casson, E. and Coles, S. (1997) Extreme hurricane wind speeds: estimation, extrapolation and spatial smoothing. In *Proc. 2nd Eur. Afr. Conf. Wind Engineering* (ed. G. Solari), vol. 1, pp. 131–138. Padova: Grafice.

Clayton, D. and Kaldor, J. (1987) Empirical Bayes estimates of age-standardized relative risks for use in disease mapping. *Biometrics*, **43**, 671–682.

Cressie, N. (1988) Spatial prediction and ordinary kriging. *Math. Geol.*, **20**, 405–421; erratum, **21** (1989), 493–494.

———(1991) *Statistics for Spatial Data*. New York: Wiley.

———(1993) *Statistics for Spatial Data*, revised edn, p. 108. New York: Wiley.

Cressie, N. and Chan, N. H. (1989) Spatial modelling of regional variables. *J. Am. Statist. Ass.*, **84**, 393–401.

De Gruijter, J. J. and ter Braak, C. J. F. (1990) Model-free estimation from spatial samples: a reappraisal of classical sampling theory. *Math. Geol.*, **22**, 407–415.

Denison, D. G. T., Mallick, B. K. and Smith, A. F. M. (1998) Bayesian MARS. *Statist. Comput.*, to be published.

Diggle, P. J., Liang, K.-Y. and Zeger, S. L. (1994) *The Analysis of Longitudinal Data*. Oxford: Clarendon.
Draper, D. (1995) Assessment and propagation of model uncertainty (with discussion). *J. R. Statist. Soc.* B, **57**, 45–97.
Fan, J. and Gijbels, I. (1996) *Local Polynomial Modelling and Its Applications*. London: Chapman and Hall.
Finney, D. J. (1962) Cumulants of truncated multinormal distributions. *J. R. Statist. Soc.* B, **24**, 535–536.
———(1963) Some properties of a distribution specified by its cumulants. *Technometrics*, **5**, 63–69.
Fitzmaurice, G. M. and Lipsitz, S. R. (1995) A model for binary time series data with serial odds ratio patterns. *Appl. Statist.*, **44**, 51–61.
Gavrikov, V. and Stoyan, D. (1995) The use of marked point processes in ecological and environmental forest studies. *Environ. Ecol. Statist.*, **2**, 331–344.
Gilks, W. R. and Wild, P. (1992) Adaptive rejection sampling for Gibbs sampling. *Appl. Statist.*, **41**, 337–348.
Goovaerts, P., Webster, R. and Dubois, J.-P. (1997) Assessing the risk of soil contamination in the Swiss Jura using indicator geostatistics. *Environ. Ecol. Statist.*, **4**, 31–48.
Green, P. J. and Silverman, B. W. (1994) *Nonparametric Regression and Generalized Linear Models: a Roughness Penalty Approach*. London: Chapman and Hall.
Guttorp, P. and Sampson, P. D. (1989) Discussion on Space–time modelling with long-memory dependence: assessing Ireland's wind power resource (by J. Haslett and A. E. Raftery). *Appl. Statist.*, **38**, 32–33.
Handcock, M. S., Meier, K. and Nychka, D. (1994) Comment on ''Kriging and splines: an empirical comparison of their predictive performance'' by G. M. Laslett. *J. Am. Statist. Ass.*, **89**, 401–403.
Handcock, M. S. and Stein, M. L. (1993) A Bayesian analysis of kriging. *Technometrics*, **35**, 403–410.
Haslett, J. and Raftery, A. E. (1989) Space–time modelling with long-memory dependence: assessing Ireland's wind power resource (with discussion). *Appl. Statist.*, **38**, 1–50.
Hastie, T. and Loader, C. (1993) Local regression: automatic kernel carpentry. *Statist. Sci.*, **8**, 120–143.
Hastie, T. J. and Tibshirani, R. J. (1990) *Generalized Additive Models*. London: Chapman and Hall.
Hay, J. (1998) Statistical modelling for non-Gaussian time series data with explanatory variables. *PhD Thesis*. Queensland University of Technology, Brisbane.
Heagerty, P. and Lele, S. R. (1998) A composite likelihood approach to binary data in space. *J. Am. Statist. Ass.*, to be published.
Holmes, C. and Mallick, B. (1997) Bayesian radial basis functions of variable dimension. *Neur. Computn*, to be published.
Høst, G. (1996) Contributions to the analysis of spatial and spatial-temporal data. *DSci Thesis*. Department of Mathematics, University of Oslo, Oslo.
Kelsall, J. E. and Wakefield, J. C. (1997) Spatial modelling of disease risk. To be published.
Kent, J. T. (1989) Continuity properties for random fields. *Ann. Probab.*, **17**, 1432–1440.
Kent, J. T. and Mardia, K. V. (1994) The link between kriging and thin plate splines. In *Probability, Statistics and Optimization: a Tribute to Peter Whittle* (ed. F. P. Kelly), pp. 325–339. Chichester: Wiley.
Kent, J. T. and Wood, A. T. A. (1997) Estimating the fractal dimension of a locally self-similar Gaussian process by using increments. *J. R. Statist. Soc.* B, **59**, 679–699.
Kitanidis, P. K. (1986) Parameter uncertainty in estimation of spatial functions: Bayesian analysis. *Wat. Resour. Res.*, **22**, 499–507.
Kolmogorov, A. N. (1941) Interpolirovanie i ekstrapolivanie statsionarnykh sluchainykh posledovatel 'nostei' (Interpolated and extrapolated random sequences). *Isv. Akad. Nauk. SSSR Ser. Mat.*, **5**, no. 1.
Krige, D. G. (1966) Two-dimensional weighted moving average trend surfaces for ore-evaluation. *J. S. Afr. Inst. Minng Metall.*, **66**, 13–38.
Laslett, G. M. (1994) Kriging and splines: an empirical comparison of their predictive performance in some applications (with discussion). *J. Am. Statist. Ass.*, **89**, 391–409.
Lawson, A. B. (1994) On using spatial Gaussian priors to model heterogeneity in environmental epidemiology. *Statistician*, **43**, 69–76.
———(1997) Some spatial statistical tools for pattern recognition. In *Quantitative Approaches in Systems Analysis* (eds A. Stein, F. W. T. P. de Vries and J. Schut), vol. 7, pp. 43–58. C. T. de Wit Graduate School for Production Ecology.
———(1998) Markov chain monte carlo methods for putative sources of hazard and general clustering. In *Disease Mapping and Risk Assessment for Public Health Decision Making* (eds A. B. Lawson, D. Boehning, E. Lesaffre, A. Biggeri, J.-F. Viel and R. Bertollini). Chichester: Wiley. To be published.
Lawson, A. B., Biggeri, A. and Lagazio, C. (1996) Modelling heterogeneity in discrete spatial data models via MAP and MCMC methods. In *Proc. 11th Int. Wrkshp Statistical Modelling* (eds A. Forcina, G. M. Marchetti, R. Hatzinger and G. Galmacci), pp. 240–250. Citta di Castello: Graphos.
Lawson, A. B. and Cressie, N. (1998) Spatial statistical methods for environmental epidemiology. In *Handbook of Statistics: Bioenvironmental and Public Health Statistics* (eds C. R. Rao and P. K. Sen). Amsterdam: North-Holland. To be published.
Longmore, M. E., O'Leary, B. M., Rose, C. W. and Chandica, A. L. (1983) Mapping soil erosion and accumulation with the fallout isotope Caesium-137. *Aust. J. Soil Res.*, **21**, 373–385.

MacKay, D. J. C. (1993) Bayesian methods for backpropagation networks. In *Models of Neural Networks II* (eds J. L. van Hemmen, E. Domany and K. Schulten). Berlin: Springer.
Matheron, G. (1965) *Les Variables Régionalisées et Leur Estimation*. Paris: Masson.
McCulloch, C. E. (1997) Maximum likelihood algorithms for generalized linear mixed models. *J. Am. Statist. Ass.*, **92**, 162–170.
Mollié, A. (1996) Bayesian mapping of disease. In *Markov Chain Monte Carlo in Practice* (eds W. R. Gilks, S. Richardson and D. J. Spiegelhalter). New York: Chapman and Hall.
Moss, R., Elston, D. A. and Watson, A. (1997) Spatial asynchrony and demographic travelling waves during red grouse population cycles. Submitted to *Ecology*.
Neal R. M. (1996) Bayesian learning for neural networks. *Lect. Notes Statist.*, **118**.
———(1997) Monte Carlo implementation of Gaussian process models for Bayesian regression and classification. *Technical Report 9702*. Department of Statistics, University of Toronto, Toronto. (Available from `http://www.cs.toronto.edu/~radford/`.)
O'Connell, M. A. and Wolfinger, R. D. (1997) Spatial regression models, response surfaces, and process optimization. *J. Comput. Graph. Statist.*, **6**, 224–241.
O'Hagan, A. (1978) Curve fitting and optimal design for prediction (with discussion). *J. R. Statist. Soc.* B, **40**, 1–42.
Oliver, M. A., Lajaunie, C., Webster, R., Mann, J. R. and Muir, K. E. (1993) Binomial cokriging the risk of a rare disease. *C. R. J. Géostatist.*, **3**, 159–165.
Omré, H. and Halvorsen, K. B. (1989) The Bayesian bridge between simple and universal kriging. *Math. Geol.*, **21**, 767–786.
Pettitt, A. N. and McBratney, A. B. (1993) Sampling designs for estimating spatial variance components. *Appl. Statist.*, **42**, 185–209.
Rivoirard, J. (1994) *Introduction to Disjunctive Kriging and Non-linear Geostatistics*. Oxford: Oxford University Press.
Sacks, J., Welch, W. J., Mitchell, T. J. and Wynn, H. P. (1989) Design and analysis of computer experiments. *Statist. Sci.*, **4**, 409–435.
Sampson, P. D. and Guttorp, P. (1992) Nonparametric estimation of nonstationary spatial covariance structure. *J. Am. Statist. Ass.*, **87**, 108–119.
She, Z.-S. (1991) Physical model of intermittency in turbulence: near-dissipation-range non-Gaussian statistics. *Phys. Rev. Lett.*, **66**, 600–603.
Stein, A. and Corsten, I. C. A. (1991) Universal kriging and cokriging as a regression procedure. *Biometrics*, **47**, 575–587.
Stein, A., Staritsky, I. G., Bouma, J., van Eijnsbergen, A. C. and Bregt, A. K. (1991) Simulation of moisture deficits and areal interpolation by universal cokriging. *Wat. Resour. Res.*, **27**, 1963–1973.
Sutherland, R. A. (1994) Spatial variability of ^{137}Cs and the influence of sampling on estimates of sediment redistribution. *Catena*, **21**, 57–71.
Tan, W. Y. (1980) On approximating multivariate distributions. In *Multivariate Statistical Analysis* (ed. R. P. Gupta), pp. 237–249. Amsterdam: North-Holland.
Tjelmeland, H. and Besag, J. (1996) Markov random fields with higher order interactions. Preprint.
Van Groenigen, J. W., Stein, A. and Zuurbier, R. (1997) Optimization of environmental sampling using interactive GIS. *Soil Technol.*, **10**, 83–98.
Venables, W. N. and Ripley, B. D. (1994) *Modern Applied Statistics with S-Plus*, 2nd edn. New York: Springer.
Wackernagel, H. (1995) *Multivariate Geostatistics*. Berlin: Springer.
Wahba, G. (1990) Spline models for observational data. *Regl Conf. Ser. Appl. Math.*
Wälder, O. and Stoyan, D. (1996) On variograms in point process statistics. *Biometr. J.*, **38**, 895–905.
Warnes, J. J. (1987) Applications of spatial statistics in petroleum geology. *PhD Thesis*. University of Strathclyde, Glasgow.
Webster, R., Oliver, M. A., Muir, K. E. and Mann, J. R. (1994) Kriging the risk of a rare disease from a register of diagnoses. *Geogr. Anal.*, **26**, 168–185.
Weir, I. and Pettitt, A. N. (1997) Binary probability maps using a hidden conditional autoregressive Gaussian process with an application to Finnish common toad data. Submitted to *Appl. Statist.*
Williams, C. K. I. and Barber, D. (1997) Bayesian classification with Gaussian processes. Submitted to *IEEE Trans. Pattn Anal. Mach. Intell.*